Sustainable Scheduling Policies for Radio Access Networks Based On LTE Technology

Ioan-Sorin Comsa

SUSTAINABLE SCHEDULING POLICIES FOR RADIO ACCESS NETWORKS BASED ON LTE TECHNOLOGY

by

IOAN-SORIN COMȘA

A thesis submitted to the University of Bedfordshire in partial fulfilment of the requirements for the degree of Doctor of Philosophy

November 2014

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate: Ioan-Sorin COMȘA Signature:

Date: November 2014

To my dear parents,

Otilia and Iulian Comşa

To my lovely niece,

Patricia Comşa

ACKNOWLEDGEMENTS

The doctoral research presented in this dissertation has been carried out mainly in the Institute of Complex Systems, Haute Ecole d'Ingénierie et d'Architecture de Fribourg (HEIA-FR), University of Applied Sciences of Western Switzerland, Fribourg, Switzerland. During these years, many people and researchers have been involved in this project and I am very grateful for their support, but in the following, there are some special persons I want to thank in particular.

First of all, I would like to thank to my dear parents, Otilia and Iulian Comşa for their love and unconditioned support. They have sacrificed their lives in order to make me become the person who I am today. I would have not achieved this without their presence in my life. Extra acknowledgements go to my brother, Sergiu Comşa, and to my sister-in-law, Alina Comşa, for making all these things possible. Nothing makes you happier than knowing that your family is so proud of you.

I would like to express the gratitude to my supervisory team from HEIA-FR and University of Bedfordshire. A special acknowledgement goes to Prof. Pierre Kuonen for sharing his intellectual potential and for the continuous moral support provided in this period. Prof. Kuonen is the key person who made this work possible at HEIA-FR and for this, I will be grateful for my entire life. I would like to thank the Director of Studies, Dr. Sijing Zhang, for his abilities of handling the heavy and tense moments, for the long-distance moral and intellectual support and for helping me to improve the thesis quality. Big thanks also go to Dr. Mehmet Emin Aydin who supported me in drawing the requirements of this project and in making the things as simple as possible. I am

very grateful to Dr. Jean-Frédéric Wagen from HEIA-FR for the very constructive and fruitful discussions that we shared together. I have benefited greatly from your academic experience. I thank you all for making me the researcher that I am today.

We all feel blessed when we are surrounded by incredible and supportive friends in our particular or professional lives. Dr. Ramona Trestian from the Middlesex University London is one of these friends who helped me a lot in my professional career for more than 10 years. There are many things for which I am very grateful, but in particular, I would like to thank Ramona for sending me the job information of this project and for making this success possible.

I have had the wonderful chance of spending the whole period of my doctoral study in one of the most beautiful countries on the planet, Switzerland. During these years, I met wonderful persons who have made great impact on my personal and professional life. I would like to give my heartfelt thanks to my colleagues Jean-François Roche and Ngoc Thuy Nguyen for their friendship and for the logistic support provided in achieving the degree. Also, I would like to express my gratitude to my office colleagues, Lu Yao and Jianping Chen. Many thanks go to the wonderful academic board from HEIA-FR: Prof. Jean Hennebert, Prof. François Killchoer, Prof. Antoine Delley, Prof. Omar Abou Khaled, Prof. Olimpia Mamula-Steiner, Christophe Schaer, Dr. Gérôme Bovet, Dr. Maurizio Caon, Cristophe Gisler, Beat Wolf, Laurent Winkler and Dr. Huang Ye.

Achieving a Ph.D. degree is a good opportunity to value your innovation capabilities and it represents a difficult task when you are far away from your family. Unfortunately, there were more difficult situations I was forced to deal with at the initial stage of my doctoral research. In this sense, I want express my gratefulness to: Pierre Kuonen, Sijing Zhang, Carsten Maple, Yong Yue, Mitul Shukla and Dayou Li. Thank you all for being my second “family” in the most difficult moments.

Fribourg, November 2014

Ioan-Sorin COMȘA

LIST OF PUBLICATIONS

Journal Paper

I.-S. Comşa, M. Aydin, S. Zhang, P. Kuonen and J. F. Wagen, "Multi Objective Resource Scheduling in LTE Networks Using Reinforcement Learning," in *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 3(2), pp. 39-57, April 2012.

Conference Papers

L. Yao, **I.-S. Comşa**, P. Kuonen and B. Hirsbrunner, "Dynamic Data Aggregation Protocol based on Multiple Objective Tree in Wireless Sensor Networks," *in IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing,* pp. 1-7, Apr. 2015.

I.-S. Comşa, S. Zhang, M. Aydin, J. Chen, P. Kuonen and J. F. Wagen, "Adaptive Proportional Fair Parameterization Based LTE Scheduling Using Continuous Actor-Critic Reinforcement Learning," in *IEEE Global Communications Conference (GLOBECOM)*, pp. 4387 - 4393, Dec. 2014.

I.-S. Comşa, M. Aydin, S. Zhang, P. Kuonen, J. F. Wagen and L. Yao, "Scheduling Policies Based on Dynamic Throughput and Fairness Tradeoff Control in LTE-A Networks," in *39th Annual IEEE Conference on Local Computer Networks (LCN)*, pp. 418-421, Sept. 2014.

J. Chen, Y. Lu, **I.-S. Comşa** and P. Kuonen, "A Scalability Hierarchical Fault Tolerance Strategy: Community Fault Tolerance,*" in 20th International Conference on Automation and Computing (ICAC)*, pp. 212 – 217, Sept. 2014.

Y.Lu, J. Chen, **I.-S. Comşa**, P. Kuonen and B. Hirsbrunner, "Construction of Data Aggregation Tree for Multi-Objectives in Wireless Sensor Network Through Jump Particle Swarm Optimization,*" in 18th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, vol. 35, pp. 73-82, Sept. 2014.

Y.Lu, J. Chen, **I.-S. Comşa** and P. Kuonen, "Backup Path with Energy Prediction based on Energy-Aware Spanning Tree in Wireless Sensor Networks,*" in International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC),* pp. 392-397*,* Oct. 2013.

I.-S. Comşa, S. Zhang, M. Aydin, P. Kuonen, and J. F. Wagen, "A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-Advanced Technologies Using Neural Networks, *"* in *37th Annual IEEE Conference on Local Computer Networks (LCN)*, pp. 332-335, Oct. 2012.

I.-S. Comşa, M. Aydin, S. Zhang, P. Kuonen and J. F. Wagen, "Reinforcement Learning based Radio Resource Scheduling in LTE-Advanced,*"* in *17th International Conference on Automation and Computing (ICAC)*, pp. 219 – 224, Sept. 2011.

SUSTAINABLE SCHEDULING POLICIES FOR RADIO ACCESS NETWORKS BASED ON LTE TECHNOLOGY

IOAN-SORIN COMȘA

ABSTRACT

In the LTE access networks, the Radio Resource Management (RRM) is one of the most important modules which is responsible for handling the overall management of radio resources. The packet scheduler is a particular sub-module which assigns the existing radio resources to each user in order to deliver the requested services in the most efficient manner. Data packets are scheduled dynamically at every Transmission Time Interval (TTI), a time window used to take the user's requests and to respond them accordingly. The scheduling procedure is conducted by using scheduling rules which select different users to be scheduled at each TTI based on some priority metrics. Various scheduling rules exist and they behave differently by balancing the scheduler performance in the direction imposed by one of the following objectives: increasing the system throughput, maintaining the user fairness, respecting the Guaranteed Bit Rate (GBR), Head of Line (HoL) packet delay, packet loss rate and queue stability requirements. Most of the static scheduling rules follow the sequential multi-objective optimization in the sense that when the first targeted objective is satisfied, then other objectives can be prioritized. When the targeted scheduling objective(s) can be satisfied at each TTI, the LTE scheduler is considered to be optimal or feasible. So, the scheduling performance depends on the exploited rule being focused on particular objectives.

This study aims to increase the percentage of feasible TTIs for a given downlink transmission by applying a mixture of scheduling rules instead of using

one discipline adopted across the entire scheduling session. Two types of optimization problems are proposed in this sense: *Dynamic Scheduling Rule based Sequential Multi-Objective Optimization* (DSR-SMOO) when the applied scheduling rules address the same objective and *Dynamic Scheduling Rule based Concurrent Multi-Objective Optimization* (DSR-CMOO) if the pool of rules addresses different scheduling objectives. The best way of solving such complex optimization problems is to adapt and to refine scheduling policies which are able to call different rules at each TTI based on the best matching scheduler conditions (states). The idea is to develop a set of non-linear functions which maps the scheduler state at each TTI in optimal distribution probabilities of selecting the best scheduling rule. Due to the multi-dimensional and continuous characteristics of the scheduler state space, the scheduling functions should be approximated. Moreover, the function approximations are learned through the interaction with the RRM environment. The Reinforcement Learning (RL) algorithms are used in this sense in order to evaluate and to refine the scheduling policies for the considered DSR-SMOO/CMOO optimization problems. The neural networks are used to train the non-linear mapping functions based on the interaction among the intelligent controller, the LTE packet scheduler and the RRM environment.

In order to enhance the convergence in the feasible state and to reduce the scheduler state space dimension, meta-heuristic approaches are used for the channel statement aggregation. Simulation results show that the proposed aggregation scheme is able to outperform other heuristic methods. When the aggregation scheme of the channel statements is exploited, the proposed DSR-SMOO/CMOO problems focusing on different objectives which are solved by using various RL approaches are able to: increase the mean percentage of feasible TTIs, minimize the number of TTIs when the RL approaches punish the actions taken TTI-by-TTI, and minimize the variation of the performance indicators when different simulations are launched in parallel. This way, the obtained scheduling policies being focused on the multi-objective criteria are *sustainable*.

Keywords: *LTE, packet scheduling, scheduling rules, multi-objective optimization, reinforcement learning, channel, aggregation, scheduling policies, sustainable*.

Table of Content

List of Figures

List of Tables

List of Abbreviations

3GPP: Third Generation Partnership Project

ACLA: Actor Critic Learning Automata

ALF: Augmented Lagrangian Function

AMC: Adaptive Modulation and Coding Scheme

ANN: Artificial Neural Network

APC: Adaptive Power Control

AS: Access Stratum

AUT-EMF: Average User Throughput with Exponential Moving Filter

AUT-MMF: Average User Throughput with Median Moving Filter

BE: Best Effort

BLER: BLock Error Rate

CACLA: Continuous Actor Critic Learning Automata

CDF: Cumulative Distribution Function

CEL: Consecutive Error Loss

CMOO: Concurrent Multi-Objective Optimization

CN: Core Network

CPU: Central Processing Unit

CQI: Channel Quality Indicator

CQI-CS: Channel Quality Indicator Classification Stage

CQI-MMR: Channel Quality Indicator Mass Mode Report

CQI-NMR: Channel Quality Indicator Normal Mode Report

CQI-PS: Channel Quality Indicator Preprocessing Stage

CQI-RS: Channel Quality Indicator Regression Stage

DCI: Downlink Control Information

DMU: Dynamic Marginal Utility

D-P: Dynamic Programming

DP: HoL Delay and Packet Drop Rate Objectives

DRL: Distributed Reinforcement Learning

DSR-CMOO: Dynamic Scheduling Rule based CMOO

DSR-SMOO: Dynamic Scheduling Rule based SMOO

eNodeB: evolved Node B

EPC: Evolved Packet Core Network

EPS: Evolved Packet System

ER: Experience Replay

E-RAB: E-UTRAN Radio Access Bearer

E-SMLC: Evolved Serving Mobile Location Centre

E-UTRAN: Evolved Universal Terrestrial Radio Access Network

FDPS: Frequency Domain Packet Scheduling

FG: Fairness and GBR Objectives

FDD: Frequency Division Duplex

FGDP: Fairness, GBR, Delay and PDR Objectives

FTP: File Transfer Protocol

GBR: Guaranteed Bit Rate

GDP: GBR, Delay and PDR Objectives

GMLC: Gateway Mobile Location Centre

GPF-BF: Generalized Proportional Fair with Barier Function

GPF-DP: Generalized Proportional Fair with Double Parameterization

GPF-EDF: Generalized Proportional Fair with Earliest Due to Date Function

GPF-EXP1: Generalized Proportional Fair with Exponential Function 1

GPF-EXP2: Generalized Proportional Fair with Exponential Function 2

GPF-LOG: Generalized Proportional Fair with Logarithmic Function

GPF-mM: Generalized Proportional Fair with Minimum/Maximum Rates

GPF-MLWDF: Generalized Proportional Fair for Modified Largest Weighted Delay First

GPF-MDU: Generalized Proportional Fair based on Maximum Delay Utility

GPF-OPLF: Generalized Proportional Fair with Opportunistic Packet Loss Fair

GPF-PLF: Generalized Proportional Fair with Packet Loss Fair

GPF-RAD: Generalized Proportional Fair with Required Activity Detection

GPF-SP: Generalized Proportional Fair with Simple Parameterization

HARQ: Hybrid Automatic ReQuest

HeNodeB-GW: Home eNodeB GateWay

HoL: Head of Line

HSDPA: High Speed Downlink Packet Access

HSPA: High Speed Packet Access

HSS: Home Subscriber Server

JFI: Jain Fairness Index

ILAMC: Inner Loop Adaptive Modulation and Coding

LTE: Long Term Evolution

LTE-A: Long Term Evolution Advanced

LTF: Long Term Fairness

MAC: Medium Access Control

MARL: Multi-Agents Reinforcement Learning

MBR: Maximum Bit Rate

MC: Monte Carlo

MCS: Modulation and Coding Scheme

MDP: Markov Decision Process

MIMO: Multiple-In Multiple-Out

MIMO-MU: Multiple-In Multiple-Out Multi-User

MLPNN: Multi-Layer Perceptron Neural Network

MME: Mobility Management Entity

MOO: Multi-Objective Optimization

MUF: Marginal Utility Function

MUSI: Marginal Utility State Informer

MUTI: Marginal Utility Type Informer

MT: Maximum Throughput

NAS: Non-Access Stratum

NGMN: Next Generation on Mobile Networks

NUT: Normalized User Throughput

OFDMA: Orthogonal Frequency Division Multiple Access

OLAMC: Outer Loop Adaptive Modulation and Coding

PDCCH: Physical Downlink Control CHannel

PDCP: Packet Data Convergence Protocol

PDR: Packet Drop Rate

PDSCH: Physical Downlink Shared CHannel

PCRF: Policy and Charging Rule Function

PDN: Packet Data Network

PF: Proportional Fair

P-GW: PDN GateWay

PHY: Physical Layer

PLR: Packet Loss Rate

PMI: Pre-coding Matrix Indicator

PS: Packet Scheduler

PUCCH: Physical Uplink Control CHannel

PUSCH: Physical Uplink Shared Channel

QCI: Quality of Service Class Identifier

QoS: Quality of Service

RAC: Radio Admission Control

RAN: Radio Access Network

RB: Resource Block

RBFNN: Radial Basis Function Neural Network

RDL: Relative Distortion Loss

RE: Resource Element

RI: Rank Indicator

RL: Reinforcement Learning

RLC: Radio Link Control

RR: Round Robin

RRC: Radio Resource Control

RRM: Radio Resource Management

SA: Simulated Annealing

SAE: System Architecture Evolution

SAST: Simulated Annealing with Stochastic Tunneling

SC-FDMA: Single Carrier Frequency Division Multiple Access

S-GW: Serving GateWay

SINR: Signal-to-Interference-and-Noise-Ratio

SMOO: Sequential Multi-Objective Optimization

SMU: Static Marginal Utility

SRS: Sounding Reference Signals

ST: Stochastic Tunneling

STD: STandard Deviation

STF: Short Term Fairness

SVM: Support Vector Networks or Support Vector Machine

TB: Transport Block

TBS: Transport Block Size

TD: Temporal Difference

TDD: Time Division Duplex

TDL: Temporal Difference Learning

TDPS: Time Domain Packet Scheduling

TTI: Transmission Time Interval

TWRG: Time Window for data Rate Guarantee

UE: User Equipment

UMTS: Universal Mobile Telecommunications System

VoIP: Voice over Internet Protocol

WCDMA: Wide-band Code Division Multiple Access

Chapter 1

Introduction

1.1 Research Motivation

The increase of mobile data usage and the growing demands for new applications (e.g., mobile television, web browsing, File Transfer Protocol (FTP), video streaming, Voice over Internet Protocol (VoIP)) have motivated 3GPP to work with LTE (3.9 Generation in Mobile Phones (G)) and LTE-Advanced (LTE-A) (4G), the latest standards of cellular communication technologies. Although the previous mobile technologies account at present for over 85% of all mobile subscribers, LTE will provide enhanced performance and benefits when compared with other technologies mentioned in [1], [2] and [11], such as: enhanced access techniques, smart antennas, spectrum efficiency and intelligent management of radio resources. From the perspective of network operators, the Radio Resource Management (RRM) includes transmission power management, mobility management, radio resource allocation and packet scheduling (PS) [1]. The packet scheduling is a process where the radio resources are assigned to each user in order to offer the requested services in an efficient way.

Based on the packet scheduling performance, the network operator is constrained in providing the requested services by using the existing radio infrastructure, regardless of the spatial/time terminal positions, user preferences, types of mobile devices or application requirements (Fig. 1.1) [3]. First of all, the

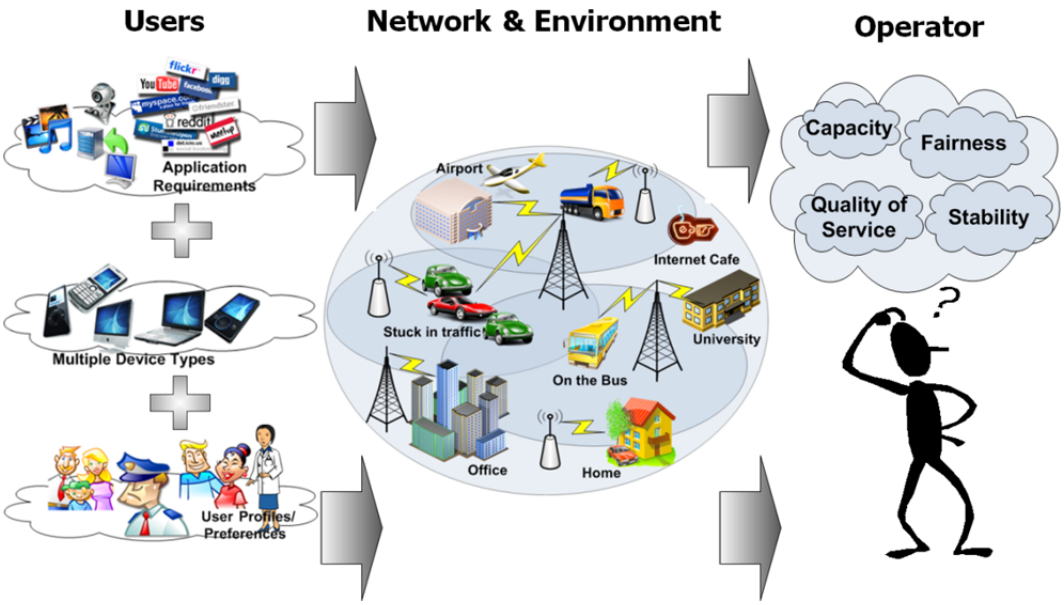


Fig. 1.1 Statement of the Scheduling Problem

main concern of the packet scheduler is to increase the system capacity or the total cell spectral efficiency, and to satisfy the application requirements for each active user at the same time. Terminals which are located near the base station experience better channel conditions and thus can receive a higher quantity of data. By providing the existing radio resources to those users with better channel conditions, other mobile terminals with poorer channel statements are starved in receiving the requested data for a longer period of time. In this sense, the fairness performance between different users with the same service profiles is strongly affected. Therefore, a proper tradeoff between the cell throughput maximization and user fairness satisfaction should be defined and maintained by the packet scheduling module.

Alongside the aforementioned aspects, the requested services should be provided in given Quality of Service (QoS) profiles such as Guaranteed Bit Rate (GBR), Head of Line (HoL) packet delay and Packet Loss Rate (PLR) requirements. The QoS requirements become more restrictive with the evolution of cellular standards, system architectures, and the application types and requirements. Another aspect which may affect the aforementioned objectives refers to the system stability. This characteristic implies in fact the maintenance of queue stability while providing the requested services under a certain QoS budget.

In other words, each data queue should guarantee the containing of enough data to be scheduled without deprecating the total cell throughput. By encompassing the discussed objectives, the packet scheduling entity should maximize the total cell throughput while maintaining a desired level of fairness among mobile users, respecting the QoS application requirements for different classes of traffic types and maintaining the queues stable during the transmission sessions on both uplink and downlink directions.

The scheduling procedure selects the user's packets based on some *priority metrics.* The packet priority metrics are calculated based on the *LTE scheduler state space information* by using a given *scheduling rule* or *scheduling discipline* for the entire transmission period. Based on the scheduling rule, each packet is scheduled in every Transmission Time Interval (TTI), a time window used to transmit the user requests and to respond them accordingly. Then the entire scheduling performance depends on the exploited scheduling rule and implicitly on the adopted performance measure.

1.2 Problem Statement

The scheduling rules are various and behave differently in the considered multi-objective satisfaction criteria. For instance, some scheduling disciplines are oriented more on the system throughput and user fairness tradeoff satisfaction by degrading the performance of QoS requirements, and some other rules are oriented on a particular QoS objective by harming the performance of other QoS objectives. Therefore, the multi-objective performance is balanced in the direction of the addressed objective being imposed by the applied scheduling rule.

The scheduler state space represents a set of observations which are involved by different scheduling rules in the metric computation at each TTI. Based on a given scheduler state at each TTI, different scheduling rules impact differently in the multi-objective performance measure. Hence, a mixture of scheduling rules can be used at each TTI instead of a single one adopted across the entire process in a way that each rule should be called based on the best matching scheduler state in order to meet the grand objective.

An efficient and successful LTE packet scheduler should be able to allocate the existing radio resources to different users in such a way that a general multi-objective satisfaction measure should be guaranteed. In this sense, the Multi-Objective Optimization (MOO) problems are addressed in order to obtain a general satisfaction level in such a way that as many particular objective measures should remain as high as possible.

The MOO model in LTE packet scheduling makes use of three main components such as: *decision variables*, *multi-objective optimization problem* and *set of constraints*. These elements influence the complexity of the multi-objective optimization model. At each TTI, the decision variables comprise the addressed objective, the scheduling rule to be applied for each user and the user selection vector for some limited number of radio resources. The proposed aggregate optimization problem represents a joint assignment of four spaces: the pool of scheduling rules, the set of objectives, the set of active users and radio resources. The objective constraints indicate the grade of satisfaction for particular objectives. Due to the product between decision variables, the MOO problem becomes non-linear. *The idea of the MOO packet scheduling is to find for each instantaneous scheduler state the best decision variables (objective, rule, user selection and radio resource assignment) in such a way that the instantaneous multi-objective optimization problem is maximized while respecting the set of objective constraints (as many as possible user objectives should be satisfied).*

Being a non-linear optimization problem, the global solution of the multi-objective approach is not guaranteed. Moreover, such MOO problems in LTE scheduling require long way of searching the best decisions for each scheduler state at each TTI. The scheduler complexity at this point is directly proportional with the number of objectives, the pool size of scheduling rules, the set size of the radio resources and the number of active users. It will be very time consuming to decide at each TTI on which objective(s) should be addressed, which scheduling rule focussing on the addressed objective(s) should be assigned and which user must be selected to transmit on a given resource. Then, other methodologies of solving such complex and dynamic optimization problems should be proposed in order to make the MOO approach suitable in real-time LTE scheduling processes.

1.3 Methodologies

In general, the problems of radio resource allocation and packet scheduling, when one single scheduling rule is applied for the entire transmission session, can be modeled by using linear programming models which convert at each TTI the scheduling problem in optimal allocation of the radio resources for a given number of active users subject to the convex set of constraints. This way, the global optimum is guaranteed. As mentioned, when the MOO considers the mixture of scheduling rules, the overall scheduling optimization problem becomes non-linear and the global optimum is not guaranteed anymore. The objective constraints can be introduced in the optimization problem by using the Augmented Lagrangian function [90]. Then, the scheduler complexity can be reduced by adopting sub-optimal LTE schedulers and by dividing the non-linear problem in two linear optimization sub-problems such that:

1. **The first linear optimization sub-problem**: selects the objective and the best scheduling rule focusing on the addressed objective based on the instantaneous scheduler state space computed at each TTI;
2. **The second linear optimization sub-problem**: based on the selected scheduling rule from the first problem, the scheduling metrics are calculated and the radio resources are optimally allocated to a predefined number of active users under a convex set of constraints.

The main difficulty is denoted by the first linear programming model which can be solved by selecting at each TTI the scheduling rule which can provide the highest multi-objective tradeoff performance. However, taking the decision at each TTI on which rule should be used is a time-consuming task, and thus real time LTE schedulers cannot be implemented. So, the scheduling policy can be used to ease the scheduling rule decision at each TTI.

The best way to optimize the mixture of scheduling rules usage for the MOO problems is to perform *the adaptation and the refinement of sustainable scheduling policies. The scheduling policy refers here to the probabilities of selecting different scheduling rules from the pool of rules for a given instantaneous scheduler state TTI-by-TTI. The sustainable term indicates the fact*

that the obtained sets of scheduling policies are optimal on the long-term purpose based on various conditions provided by the PS and RRM entities. The sustainable scheduling policies can be adapted and refined at each TTI by using dynamic programming, temporal difference learning and Markov Decision Processes (MDP) in order to learn the long-term optimal policies which can be applied to each scheduler state condition.

The *scheduling policy improvement and evaluation* are performed based on the scheduler state space. The biggest disadvantage refers to the fact that the scheduler state space is *continuous* and *multi-dimensional*. This means that the policy refinement cannot be performed based on discrete scheduler state spaces, and then, function approximations are preferred in this sense to map the continuous state space at each TTI in the optimal scheduling rule selection. The problem now is to define the methodologies which can evaluate and refine (or to improve) the scheduling policies at each TTI under continuous and multi-dimensional scheduler state spaces.

Under continuous MOO tasks, the scheduler state space contains some irrelevant information which considerably increases the LTE scheduler state space dimension. By increasing the state space dimension, the function approximation calculation becomes very time consuming. In this sense, the methodologies of aggregating the LTE scheduler state space have to be defined. Then, the policy refinement is achieved based on the aggregate scheduler state space.

1.3.1 LTE Scheduler State Space Aggregation

The scheduler state space contains different performance indicators, channel statement information and system stability parameters. So, the scheduler state depends heavily on the number of active bearers since the aforementioned indicators are calculated for each active user. Therefore, the dimension of the scheduler sub-space for the performance indicators can be reduced by using statistical functions. More sophisticated models are needed for the channel statement aggregation. At the LTE base station level, the channel statement is received from each user in the form of data vector depending on the system

bandwidth known as Channel Quality Indicator (CQI). The CQI vector for each user contains discrete numbers from 1 (the worst channel condition) to 15 (the best channel condition) for each resource block. Before statistical models are used, a classification stage needs to be performed in order to classify the CQI vector for each active user in different patterns. The classification stage is performed by using two steps:

1. **Unsupervised Learning Step**: The received channel statements are grouped in different clusters (based on CQI centers). The methodology used in this sense is called k-means clustering [181-189], [196].
2. **Supervised Learning Step**: This is required in order to classify the CQI information in different CQI patterns by using approximations. The approximation is achieved through non-linear functions which are trained based on the obtained CQI data centers from the unsupervised learning step. The methodology which is used in this sense is entitled the Radial Basis Function Neural Network (RBFNN) generalization [165-167].

1.3.2 Policy Evaluation and Policy Improvement

After performing the state space aggregation procedure, the main task is to improve and to evaluate the policy of scheduling rules in order to find the generic and the most representative one which can be applied at each TTI. Even if the aggregation stage is performed, the state space remains continuous and multi-dimensional. The idea is to develop a generalization function which can directly approximate the aggregate state space in the most representative scheduling rule. The generalization function has to be learned for each scheduling rule. At each TTI, the discipline which maximizes the learned function based on the aggregate scheduler state is selected for the radio resource allocation procedure. The methodology used in this sense is called the *Multi-Layer Perceptron Neural Network* (MLPNN) function generalization [201], [202].

When the scheduling rules are approximated and applied TTI-by-TTI, the RRM environment evaluates the performed actions by providing the reward values. These are the results of the multi-objective tradeoff evaluation when

applying different scheduling rules in different scheduler states. The MLPNN functions are learned based on the interaction between the intelligent entity called *controller* and the LTE scheduler. This interaction is modeled by using MDP principles based on the aggregate scheduler state space, the applied scheduling rule and the obtained reward value. Then, the MLPNN weights are trained TTI-by-TTI by interacting with the LTE scheduler and RRM environment under the form of Temporal Difference Learning (TDL) [203]. The Reinforcement Learning (RL) [203] as a type of TDL is used to train the MLPNN weights by reinforcing the reward values at each TTI based on the considered MDP problems. The reward value can be reinforced under different functions. Basically, the type of the reinforcement function determines the type of the RL algorithm.

The RL principle models the interaction with the LTE scheduler by using two stages: *exploration* and *exploitation*. In the exploration stage, the MLPNN weights are updated based on the type of reinforcement which is used and based on the reward value. The scheduling rule is selected at each TTI based on two principles:

1. **Policy Evaluation**: The selected scheduling rule maximizes the MLPNN functions based on the current aggregate scheduler state space.
2. **Policy Improvement**: The selected scheduling rule can be selected randomly and can be different from the rule provided from the scheduling policy trained so far.

The policy evaluation and policy improvement can be switched during the exploration period based on some probability distributions. The exploitation stage evaluates the learned policy and the trained MLPNN functions (and the scheduler reward is not reinforced anymore in the MLPNN structures). Due to the over-fitting or local optimum problems which exist in the MLPNN generalizations, the *experience replay stage* may be introduced between exploration and exploitation stages in order to avoid the aforementioned problems. *If the generalization of the non-linear MLPNN functions can provide optimal scheduling rules for various conditions of the scheduler state space in the long-term purpose, then the exploited scheduling policy becomes sustainable.*

1.4 Aims and Objectives

Based on some forecast studies which reveal the dominance of the downlink traffic type in LTE/LTE-A systems [199], the proposed scheduling policies use a mixture of scheduling rules which are performed only in the downlink transmission sense. The pool of scheduling rules which is used is focusing on one or multiple scheduling objectives. The general objective of this research is to solve a given scheduling optimization problem in order to increase the number of feasible TTIs when different performance criteria are addressed. The sustainable scheduling policies are learned based on the aggregate scheduler state space. The objectives of the proposed aggregation technique for the scheduler state space are listed below.

1.4.1 LTE Scheduler State Space Aggregation

The most important processing unit in the LTE state space aggregation is the classification procedure. As mentioned above, the classification procedure includes the unsupervised and supervised learning steps. During the unsupervised learning step, the data centers for the CQI statements are determined. In this case, the main objective is to propose a novel clustering method which is able *to minimize the squared-error distortion* between the obtained set of CQI data centers and each existing CQI vector from the data set. In the supervised learning step, the idea is to use the RBFNN function approximation in order to classify the channel quality vector in desired patterns. In this sense, the objective is to train the RBFNN weights in order to *minimize the mean squared error* between the RBFNN outputs and the given patterns. The classified observations of CQI statements can be used by the additional regression stage, in which only the most relevant characteristics of the channel statistics will be used in forming the scheduling policies. For instance, the current study proposes to use some statistical models in order to reduce much more the CQI state space dimension without losing the integrity of the provided information. The regressed observations can be used in the aggregate state space to approximate, through MLPNN functions, a given scheduling rule based on different RL techniques.

1.4.2 Sustainable Scheduling Policies Based Multi-Objective Optimization

First of all, an optimization problem is required in which different scheduling rules addressing particular objectives should be included in the mathematical model. As mentioned earlier, in this case of multi-objective optimization, the packet scheduling procedure becomes a non-linear programming problem which can be divided in two linear sub-problems in order to speed-up the scheduling process. Basically, the scheduler becomes sub-optimal and divides the scheduling problem into two stages: in the first stage, a particular scheduling rule is selected, and in the second stage, the selected rule contributes in calculating the metrics which are used in allocating the radio resources.

As discussed above, each scheduling rule addresses a particular scheduling objective such as system throughput maximization, user fairness satisfaction, QoS requirements or stability condition satisfaction. In general, one scheduling rule is focused first on the main objective; once the main objective is satisfied, the static scheduling rule can optimize other objectives with the amendment that the first objective is always satisfied. In this case, the optimization problem becomes **Sequential MOO (SMOO)**. In other circumstances, one static scheduling can optimize multiple objectives at the same time when applied TTI-by-TTI. The optimization procedure is called **Concurrent MOO (CMOO).** Based on the current proposals, the scheduling process can be divided into two main directions when the mixture of rules is used instead of one single rule being applied for the entire scheduling procedure:

1. **Dynamic Scheduling Rule based SMOO (DSR-SMOO)**: A multitude of scheduling rules which are oriented on the same objective can be used in order to enhance the scheduler convergence to the desired state from the viewpoint of the addressed objective.
2. **Dynamic Scheduling Rule based CMOO (DSR-CMOO):** A mixture of scheduling rules which are focused on different scheduling objectives can be applied at each TTI in order to enhance the scheduler convergence in the optimal state from the viewpoint of the multi-objective performance

criteria. In this sense, some performance metrics should be defined in order to maximize the number of TTIs when the scheduler is satisfied from the viewpoint of the considered objectives.

Regardless the proposed techniques of DSR-SMOO or DSR-CMOO problems, the optimal state should be reached at each TTI when a specific MOO problem is considered. The first major objective is to maximize the number of TTIs when the scheduler state is optimal from the viewpoint of different multi-objective criteria. Then, the second objective of the current study is to refine the set of scheduling policies in order to maximize, in the exploitation period, the number of TTIs when the scheduler reward is maximized. When the scheduler reward is maximized, the scheduler state becomes optimal, and implicitly the considered scheduling objectives are satisfied. On the other side, the number of punishment rewards should be minimized. Then, *the obtained set of scheduling policies becomes sustainable in the exploitation stage for various LTE scheduler conditions if the number of feasible TTIs is maximized and if the number of punishment rewards is minimized.*

1.5 Thesis Contributions

The main contributions of this research are listed below:

- The integration of DSR-SMOO/CMOO problems and the scheduler state space aggregation module in the protocol architecture of LTE/LTE-A.
- Three types of linearization techniques for the DSR-SMOO/CMOO non-linear programming problems.
- A novel classification of LTE scheduling methods based on different SSR/DSR-SMOO/CMOO approaches.
- A comprehensive survey in packet scheduling by using the novel classification scheme.
- The LTE scheduler state space aggregation: an innovative preprocessing block which is able to classify the large data vector dimension for CQI statements in predefined patterns in order to improve the quality of the results for the sustainable scheduling policies.

- A novel scheduling architecture in which the packet scheduler and the RRM environment interact with the intelligent *controller* in order to refine the sustainable scheduling policies for the DSR-SMOO/CMOO problems.
- A novel scheduling rule oriented on the GBR objective which can outperform other existing scheduling techniques for some particular traffic types.
- The reward functions for the DSR-SMOO problems focusing on user fairness criterion and focusing on GBR objective.
- The set of sustainable scheduling policies for DSR-SMOO problems being oriented on user fairness criterion and GBR objective.
- The reward functions for the DSR-CMOO problem focusing on HoL delay and Packet Drop Rate (PDR) multi-objective criterion and for the DSR-CMOO problem focusing on fairness, GBR, HoL delay and PDR objectives.
- The set of DSR-CMOO sustainable scheduling policies oriented on user fairness, GBR, PDR and HoL packet delay multi-objective criterion.
- The implementation in C/C++ language of the LTE-A-Scheduler simulator.

The simulation results of the proposed scheduling policies are conducted by using the LTE-A-Scheduler, simulator which was implemented by using the existing infrastructure being offered by the Institute of Complex Systems, University of Applied Sciences of Western Switzerland. The LTE-A-Scheduler uses the simulation model of LTE-Sim [156] by importing the radio channel models and other additional functions of the LTE protocol stack.

1.6 Thesis Outline

The thesis is organised in eight chapters as follows:

- **Chapter 1** states the LTE scheduling problem, describes the research problem and presents the methodologies which are used for the scheduler state space aggregation and for the improvement and the evaluation of the sustainable scheduling policies. The aims and objectives are addressed in order to prove the sustainability of the learned scheduling policies.
- **Chapter 2** highlights the importance of adopting the novel architecture in order to enhance the quality of the scheduling procedure in LTE/LTE-A

standards. First of all, the evolution of cellular standards, goals, requirements and the general architecture of LTE/LTE-A are discussed. The functionalities of RRM entities are analyzed and novel packet scheduler architectures are highlighted based on time domain and frequency domain scheduling correlated with SSR/DSR-SMOO/CMOO approaches. Finally, the integration of DSR-SMOO/CMOO approaches in the coupled time-frequency packet scheduling is analyzed.

- **Chapter 3** proposes linear programming models for the DSR-SMOO/CMOO MDP problems. In order to reduce the computational complexity, sub-optimal schedulers are preferred for the DSR-SMOO/CMOO problems. The proposed optimization problem includes the aggregate multi-objective constraints by using the Augmented Lagrangian function. The obtained non-linear programming problem is divided in two stages: in the first stage, the best scheduling rule which maximizes the aggregate multi-objective function and the Lagrange multiplier is selected, and in the second stage, the radio resources are optimally allocated to the active users based on the selected discipline. The first linear optimization problem is solved by modeling the RRM environment as MDP processes. The Lagrange multiplier is replaced by the accumulated reward for a given policy and the aggregate multi-objective function becomes the instantaneous reward. The RL approach is selected to be performed at each TTI in order to refine and to adapt the scheduling policies for each given continuous and multi-dimensional scheduler state space. A novel classification scheme for scheduling techniques in LTE is proposed based on the DSR-SMOO/DSR-CMOO methodologies. The related studies in LTE scheduling are analyzed based on the proposed classification scheme.
- **Chapter 4** proposes a model for the LTE scheduler state space aggregation in order to enhance the convergence to the optimal state when the RL approach is used. The CQI state space is one of the most important scheduler subspace being able to control the tradeoff between system throughput maximization and user fairness satisfaction. Due to its high dimensionality, the channel feedbacks have to be preprocessed, classified and regressed in order to avoid the dependency on the number of active users and on the system bandwidth.

The preprocessing stage reduces the dimension of each channel feedback. The classification stage is based on the unsupervised and supervised learning steps. During the unsupervised step, different sets of centers for preprocessed CQIs are obtained by using the novel meta-heuristic approach being entitled: Simulated Annealing with Stochastic Tunneling (SAST). The preprocessed CQI inputs are classified into different patterns based on the supervised step when the sets of weights for the RBFNN structures are optimized based on the provided centers. When the sets of centers and weights are trained enough, then the entire classification stage is exploited. Different statistical models are applied in the classified CQI space in order to extract the most relevant features. The simulation results show the advantage of using the proposed meta-heuristic techniques when compared with other methodologies.

- **Chapter 5** introduces the elements of interfacing the LTE scheduler and the scheduler controller which are used to learn the sustainable scheduling policies in order to solve the first linear optimization sub-problem introduced in *Chapter 3*. The principles of modeling the LTE scheduler behavior as MDP processes are discussed in order to refine the scheduling policies based on the RL methods. The analyzed RL algorithms aim to exploit the state and state-action values and to update these values at each TTI. Based on the continuous and multi-dimensional characteristics of the aggregate LTE scheduler state space and the action space (in some conditions), the MLPNN function approximations are used to map each aggregate scheduler state from *Chapter 4* in optimal scheduling rules. The implementations of various RL approaches are introduced in order to learn the best scheduling policy for a given MOO problem. Based on approximated RL approaches, the proposed architecture from *Chapter 2* is extended for multi-agent systems with specific cooperation when the entire set of scheduling objectives is taken into consideration.
- **Chapter 6** provides the sets of sustainable scheduling policies when the DSR-SMOO problems are considered being focused on fairness-system throughput tradeoff and GBR objective. In this sense, the reward functions are proposed in order to learn the optimal policies focusing on the considered objectives. The way how the average user throughput is computed plays a crucial role in

the satisfaction of GBR and fairness-throughput objectives. When the exponential filter is used, it is shown through extensive simulations that only when the aggregation schemes from *Chapter 4* are used, the proposed scheduling policies can outperform the existing techniques from the fairness requirement point of view. When the median moving filter is performed for the average throughput computation, sustainable scheduling policies being focused on fairness criterion are obtained for different window lengths of the median filter. When the performance of GBR objective is analyzed, the proposed set of scheduling policies outperforms the particular scheduling rules for different types of traffic. For the infinite buffer traffic, the proposed GBR static scheduling rule focusing on GBR requirements is the best option.

- **Chapter 7** analyzes the performance of DSR-CMOO scheduling policies being oriented on different combinations of scheduling objectives such as fairness, GBR, PDR and HoL delay requirements. The reward functions and the input states are proposed for different combinations of DSR-CMOO problems. Based on the RL approaches, the obtained sustainable policies being focused on multi-objective targets are able to perform much better than standard scheduling rules focusing on particular objectives by maximizing at the same time the number of TTIs when the scheduler is declared optimal and by minimizing the amount of punishment rewards.
- **Chapter 8** concludes this research by describing the advantages of using the current contributions when compared with the existing methodologies. The limitations of the proposed approach are discussed in terms of the trade-off between exploration and exploitation when the RL methodology is used in LTE scheduling. The possible hardware implementations are also presented in order to prove the eligibility of the proposed set of scheduling policies. Some aspects of the future research directions are highlighted in terms of the packet scheduling based on the multi-class traffic types.

This work is accompanied by the auxiliary material which is organized as follows:

- **Appendix A** presents an extended overview of the related work on SSR-SMOO and SSR-CMOO problems focusing on the following objectives such as: system throughput, user fairness, GBR, HoL delay and queue stability.

- **Appendix B** introduces the CQI cycle in LTE networks with edifying results that show the importance of the fading models in the SINR levels. Then, the quantization procedure from the SINR to the discrete CQI vector is presented.
- **Appendix C** proposes an innovative preprocessing model of the CQI reports which is absolutely mandatory in the CQI state space aggregation in order to eliminate the bandwidth dependency of the CQI state space. This appendix is an extension of *Chapter 4*.
- **Appendix D** presents an extended set of simulation results for different clustering algorithms with different system bandwidths and number of CQI data centers. The proposed SAST heuristic algorithm for k-means clustering shows the best performance when compared against the traditional approaches from the viewpoint of the best average distortion.
- **Appendix E** extends the set of simulation results from *Chapter 4* for the traditional heuristic algorithms in k-means clustering. The performances are analyzed in terms of the best average distortion and the computation complexity when the number of CQI data centers varies in a large domain for each CQI data collection with different LTE bandwidths.
- **Appendix F** evaluates the performance of sustainable scheduling policies being oriented on NGMN fairness criterion. The simulation results are conducted through various RL scheduling policies and evaluated in terms of mean percentage of feasible TTIs and mean percentage of TTIs for different types of testing rewards. This appendix is the supplemental material for the simulation results which are provided in *Chapter 6*.
- **Appendix G** analyzes the performance of sustainable scheduling policies focusing on the GBR objective from the viewpoint of mean percentage of GBR feasible TTIs and the number of different types of testing rewards. This appendix is an extension of *Chapter 6*.
- **Appendix H** evaluates the performance of scheduling policies focusing on HoL packet delay and PDR multi-objective criterion. This work extends the simulation results from *Chapter 7*.

Chapter 2

LTE Packet Scheduling: Background and Preliminaries

2.1 Chapter Outline

This chapter addresses the main principles that underlie the LTE packet scheduling procedure. The aggressive demands for QoS requirements impose higher dynamicity of the LTE packet scheduler. The proposed LTE scheduling concept is able to eliminate the disadvantages of the previous methodologies by improving the system performance when the scheduling objectives are taken into account. The novel architecture makes use of the intelligent controller which adapts the decisions of adopted scheduling rules based on various conditions of the LTE scheduler state space.

2.2 The Evolution of Cellular Standards

The commercial deployment evolution of the cellular standards is presented in Fig. 2.1. LTE was introduced by 3GPP in 2008 and is able to provide enhanced performances in terms of data rates and end-to-end delay through the simplified Core Network (CN) architecture entitled Evolved Packet CN (EPC) and Radio Access Network (RAN) architecture, known as Evolved Universal

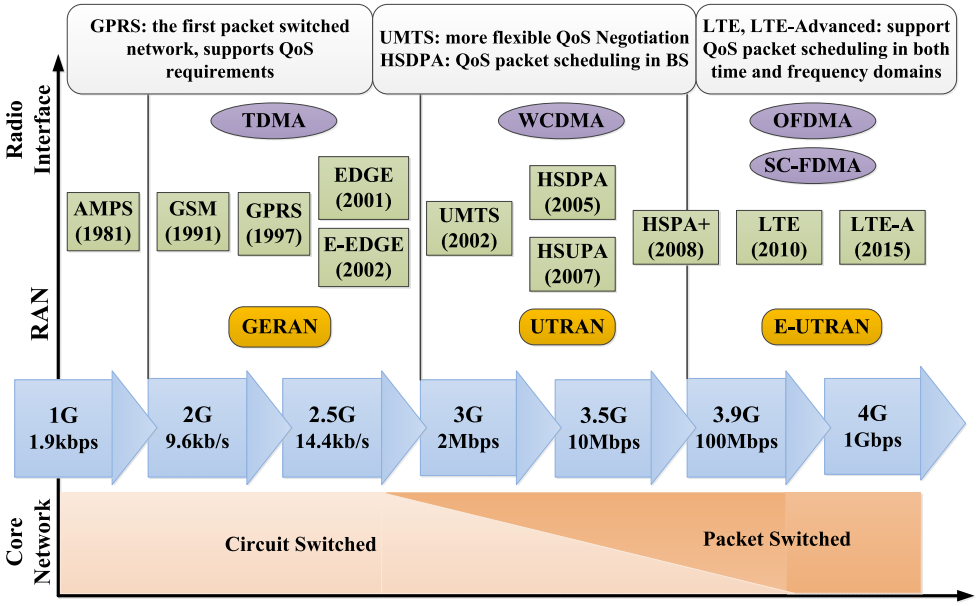


Fig. 2.1 Commercial Deployments of Cellular Networks: Past, Present and Future (based on [1-7], [41])

Terrestrial Radio Access Network (E-UTRAN). When compared with the previous standards, the radio functions in LTE are managed based on a single node entitled E-UTRAN Node-B (eNode-B). In LTE, a novel radio interface was introduced such as the Orthogonal Frequency Division Multiple Access (OFDMA) for the downlink transmission and the Single Carrier Frequency Division Multiple Access (SC-FDMA) for the uplink case. Based on the novel access techniques, receivers are simplified, the interference between neighboring cells is reduced, and different system bandwidths can be used such as 1.4MHz, 3MHz, 5MHz, 10MHz, 15MHz and 20MHz. The channel conditions of each User Equipment (UE) are transmitted on the uplink direction by using the CQI reports which in facts permit to adapt the transmission parameters based on the Adaptive Modulation and Coding (AMC) scheme. The QoS management is handled at the Evolved Node B (eNodeB) level. When compared with previous standards where the scheduling procedure is achieved in the time domain, LTE offers the possibility of scheduling mobile users in both frequency and time domains. The OFDMA based scheduling depending on the CQI measurements and QoS profiles constitute the key point of increasing the system capacity while respecting the user fairness criterion and QoS requirements.

In Europe, LTE was first launched in Norway and Sweden in December 2009 offering poor coverage and devices compatibility [8]. In Switzerland, Swisscom is the first operator who launched 4G/LTE in November 2012 [9]. Other existing operators followed the initiative of Swisscom and at the beginning of 2014 more than 70% of mobile subscribers enjoyed the LTE services. In the United Kingdom, the Everything Everywhere (EE) Limited operator launched first the LTE services in October 2012 [10]. The market predictions achieved in [10] indicate that the High Speed Packet Access (HSPA) device variants will decrease from 30% in 2012 to 7% in 2016 on favor of the LTE devices.

The main focus of this thesis is concentrated on the LTE/LTE-Advanced packet scheduler functionality on the downlink purpose which is located in the eNodeB architecture.

2.3 Goals and Requirements in LTE/LTE-A

The principal goals and requirements imposed by the LTE standard are defined in [11] and further discussed in [12], such as reduced latencies in terms of connection establishment and data transmission, peak user rates in uplink/ downlink of about 50/100 Mbps, reduced power consumption on UE terminals, enhanced mobility and security and improved cell spectral efficiency. Additional operator requirements are defined by the Next Generation Mobile Networks (NGMN) alliance according to [13]. Based on NGMN specifications, LTE/LTE-A developments are designed in order to achieve some objectives, such as user fairness requirement which is largely analyzed in Chapter 6. By using OFDMA access and multi-antenna techniques, LTE is able to provide data rates very close to the Shannon capacity [12]. Then, the main research direction of LTE-A is driven on the procedures of improving the Signal-to-Interference-and-Noise-Ratio (SINR) for a much larger set of subscribers with novel techniques such as bandwidth aggregation and support for heterogeneous architectures. In this sense, the LTE-A requirements are based on the following elements including the high level objectives [12]: improved peak data rates for uplink/downlink in order to support advanced services (100Mbps/1Gbps), worldwide roaming capability,

compatibility of services within International Mobile Telecommunications (IMT) and with fixed networks, cost-efficient, support for multi-antenna techniques and high quality services.

2.4 LTE/LTE-A System Architecture

In 3GPP Release 8 are introduced the first components and requirements for Evolved Packet System (EPS) that represents the overall system architecture for the next generation networks [14]. Two main components are considered in the original specifications: *LTE* which refers to the evolution of the radio access based on E-UTRAN and System Architecture Evolution (SAE) that comprises the evolution of non-radio functionalities including the EPC architecture. The 4G specifications bring modifications only for E-UTRAN and radio access techniques, without major changes for the EPC architecture, and details were exposed in 3GPP Releases 8 to 13 [14-19].

Every element in the EPS architecture has its own predefined role. The EPS architecture contains the core network (EPC) and the radio access (E-UTRAN), as suggested in Fig. 2.2. These components communicate with each

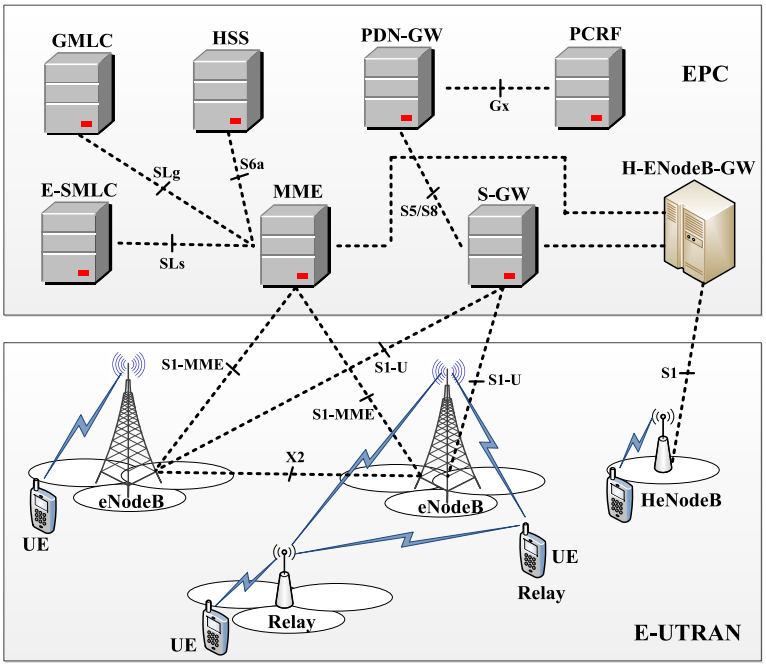


Fig. 2.2 The EPS Network Architecture

other by using the standardized interfaces allowing in this way the multivendor interoperability [11]. The IP traffic which is routed by the EPS from any external Packet Data Network (PDN) to an UE is associated with the EPS bearer. A bearer represents an IP data flow that has a specific class of QoS requirements. Basically, EPC is responsible of controlling UEs and establishing the bearers. The main components for the EPC architecture are briefly explained below:

- **Policy and Charging Rule Function (PCRF)** is responsible for QoS authorization and decides how a certain data flow will be treated based on the subscriber profile;
- **Home Subscriber Server (HSS)** contains the data subscription for users and the information about the roaming restrictions;
- **Gateway Mobile Location Centre (GMLC)**. As the name suggests, it contains the information about the estimation of the UE final location;
- **PDN GateWay (P-GW)** allocates IP addresses for UEs, filters downlink IP packets into different bearers and provides mobility functionalities when inter-working with non-3GPP networks.
- **Serving Gateway (S-GW)** serves as a mobility anchor when an UE moves between eNodeBs or between different 3GPP RANs. The component also, helps achieve administrative functions such as retaining information about UE in the IDLE state, collecting volume of information sent/received to/from UE, and legal interception.
- **Mobility Management Entity (MME)** assures the signaling processes between UE and CN based on the Non Access Stratum (NAS) protocol; supports a set of functions such as bearer management, connection and mobility management and inter-working with other networks.
- **Evolved Serving Mobile Location Centre (E-SMLC)** estimates the UE final location and the UE speed.
- **Home eNodeB-GW (HeNodeB-GW)** manages several thousands of HeNodeBs from the MME perspective.

As said, one of the major improvements of E-UTRAN is the elimination of the centralized controller in favor of a distributed or flat architecture. Moreover,

each eNodeB has its own controller. By using such a distributed control, it eliminates the need of a high processing controller and reduces the latencies, improves the efficiency and avoids the data loss during the handover procedures. As shown by Fig. 2.2, E-UTRAN consists of the eNodeBs, relays, HeNodeBs, UE equipments and the associated interfaces [14]. The most important E-UTRAN functions include Radio Resource Management (RRM), security, packet header compression and UE positioning. The RRM entity covers the uplink and downlink functionalities regarded to the radio bearer control, Radio Admission Control (RAC), mobility of radio bearers, packet scheduling and dynamic radio resources allocation. Because the scheduling process in LTE/LTE-Advanced is located at the eNodeB base station level, the main attention of this study will be focused on the E-UTRAN radio access architecture.

2.5 Quality of Service in LTE/LTE-A

Even if many aspects of QoS from Universal Mobile Telecommunications System (UMTS) can be applied to LTE, several changes are done due to the flat infrastructure of LTE (reduced number of processing units) [21]. By reducing the number of the networking units, the LTE/LTE-A QoS requirements become more restrictive as described in [23]. These aspects affect the scheduling performance as the QoS requirements for a traffic category have to be satisfied simultaneously.

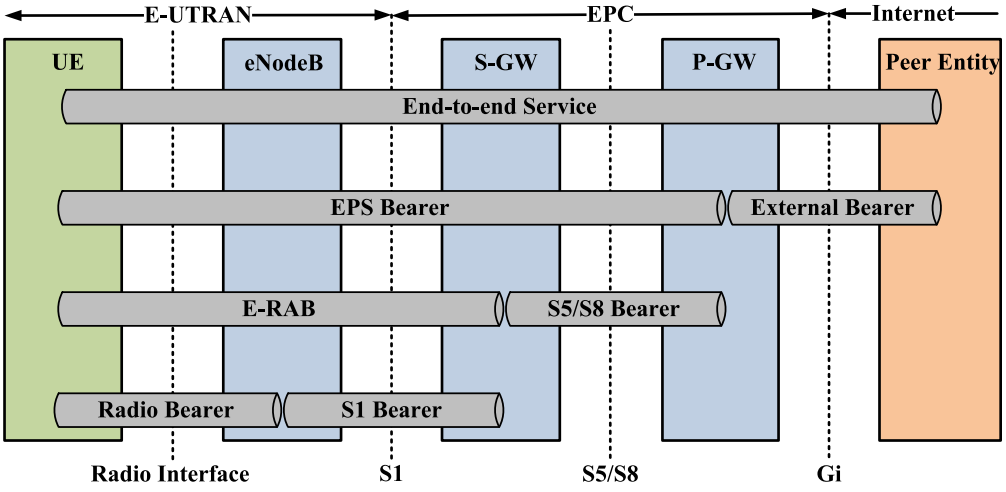


Fig. 2.3 The EPS Bearer Architecture (reproduced from [24])

Broadly speaking, the QoS concept is regarded to the network capacity to offer end-to-end service by satisfying an imposed level of service [22]. The service level refers to a set of objectives or QoS parameters that should be respected for different service types in order to satisfy and to improve the user experience. The QoS service level is determined based on the negotiation with the PCRF entity. The way how the service level is respected for each terminal strongly depends on the performance of the packet scheduling scheme located at the eNodeB level. This study proposes a novel scheduling technique that is able to increase the user satisfaction in terms of different objectives when compared with other existing techniques.

The set of QoS objectives are represented by the concept of bearer which represents in fact a logical connection between two nodes as shown in Fig. 2.3 [24]. The EPS bearer has to cross multiple entities until it reaches the final UE destination entity. The EPS radio bearer is represented by the tunneling protocol between P-GW and UE. The E-UTRAN Radio Access Bearer (E-RAB) is used to map a S1 bearer into a radio bearer. The scheduling procedure takes into consideration the radio bearers which are established between eNodeB and multiple UE entities. Based on the service type and on the number of active applications, multiple radio bearers can be defined per each UE. Basically, the radio bearers can be divided into two main categories [23]:

- **GBR**. Dedicated radio resources are required in order to achieve a minimum bit rate. The surplus of bit rates that exceed the GBR level can be upper limited by using the Maximum Bit Rate (MBR) requirement. It is important to note that by imposing the MBR bound, some resources may be conserved in order to allocate more resources for those bearers that do not meet the GBR objective.
- **Non-GBR.** There is not any requirement on the minimum bit rate, and thus, the radio resources could be allocated based on different performance criteria (e.g. HoL packet delay, packet drop/loss rate).

The packet scheduler entity from eNodeB is responsible of achieving the QoS requirements for each radio bearer. The QoS objectives are identified in LTE

Table 2.1 The Standardized QoS Class Identifier for LTE (reproduced from [23])

QCI	Resource Type	Priority	Packet Delay Budget [ms]	Packet Loss Rate	Service Examples
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	5	300	10^{-6}	Non-Conversational video (buffered streaming)
4	GBR	3	50	10^{-3}	Real time gaming
5	Non-GBR	1	100	10^{-6}	IMS signaling
6	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
7	Non-GBR	6	300	10^{-6}	Video (buffered streaming)
8	Non-GBR	8	300	10^{-6}	TCP-based (WWW, e-mail)
9	Non-GBR	9	300	10^{-6}	FTP, chat, p2p file sharing, progressive video, etc.

based on the standardized QoS Class Identifier (QCI) which is in charge of allocating different performance targets depending on the traffic type (Table 2.1). The QCI classes are represented by resource type, priority, packet delay budget (PDB) and packet loss rate (PLR) [23].

2.6 The LTE Protocol Architecture

The eNodeB base station offers for the E-UTRAN radio interface both the user plane and control plane protocol stacks. Figure 2.4 provides a brief overview of these protocols. For the user plane (U-plane) part, four protocols are included, i.e., Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), *Medium Access Control* (MAC) and Physical Layer (PHY). The control plane (C-Plane) adds the NAS procedures and the Radio Resource Control (RRC) as the Access Stratum (AS) control protocol. The important features of these protocols are presented in the following sub-sections, and the details can be found in selected bibliography [25-30].

The MAC functionalities can be divided in three categories [29-34]: dedicated eNodeB-UE MAC, dedicated UE and dedicated eNodeB MAC functions. The dedicated eNodeB-UE functions comprise the multiplexing/de-multiplexing and the mapping procedures from logical to transport channels, Hybrid Automatic ReQuest (HARQ) and the Transport Block (TB) computation.

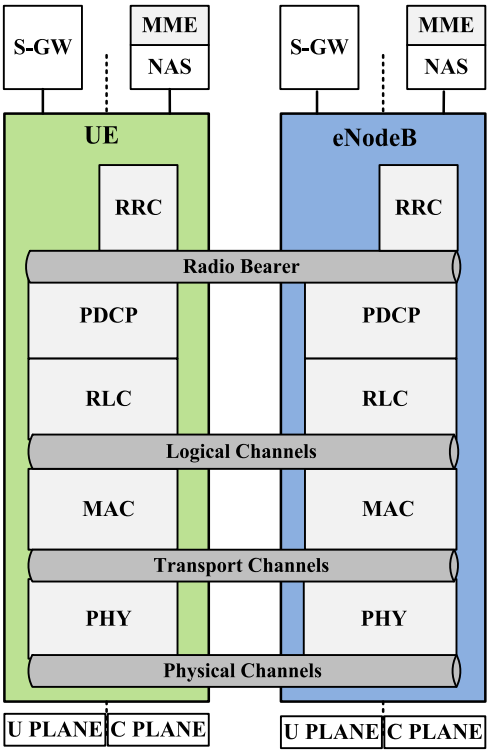


Fig. 2.4 The LTE Protocol Stack

The dedicated UE MAC procedures include the terminal energy saving procedures, scheduling request messages and CQI report information [29-33]. The dedicated eNodeB MAC procedures include the packet scheduling entity and the TB size computation as a result of the scheduling process [29]. The packet scheduler placed in the eNodeB MAC layer distributes the available radio resource of one cell to different UEs or to different radio bearers defined for each UE. From the implementation point of view, the scheduling function is defined in two ways: the scheduling algorithm and the signaling associated with the scheduling framework. If the signaling procedures are clearly standardized by 3GPP, the scheduling algorithm is left on the desire of the mobile operators. The lack of the scheduling algorithm standardization causes the controversy in the research communities due to the fact that each algorithm has a different impact in the overall system performance. The current research tries to eliminate this controversy by exploiting the particularity of each analyzed scheduling rule.

The LTE PHY layer is designed in order to offer enhanced radio access techniques in terms of OFDMA/SC-FDMA by providing at the same time the compatibility with the previous radio access techniques [34]. The technical

specifications for PHY layer functionalities in both directions can be found in [30]. For the downlink transmission, OFDMA provides higher scalability, simpler equalization techniques and higher robustness in the frequency domain when compared with Wide-band Code Division Multiple Access (WCDMA) [20].

2.7 Resource Scheduling in OFDMA

The OFDMA technique provides the possibility of allocating users by using a frequency sub-carrier of 15KHz and a symbol duration. Therefore, the allocation methodology considers the time and frequency domains. In WCDMA, users are allocated in the time domain by using the entire frequency bandwidth with different spreading codes. By having the possibility of scheduling users on different sub-carriers, a new concept is introduced in terms of frequency diversity. The frequency diversity implies the Frequency Domain Packet Scheduling (FDPS). By increasing the number of users in the FDPS scheduling, the multiuser diversity can be improved. Therefore, by exploiting the multiuser diversity, the FDPS scheduling procedure is able to increase the overall cell capacity. By scheduling users based on sub-carrier granularity implies a higher signaling overhead. LTE avoids this drawback by grouping 12 sub-carriers into Resource Block (RB) representation (Fig. 2.5). The RB corresponds to the smallest resource

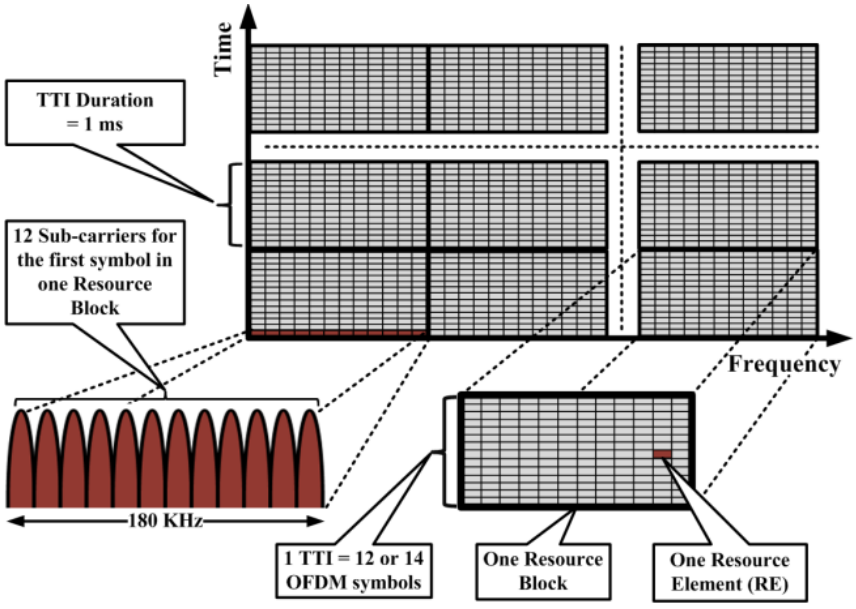


Fig. 2.5 Conceptual Resource Allocation in LTE

unit that can be assigned for one UE in the FDPS process. The RB granularity spans on 180 KHz in the frequency domain and on 1ms (12 or 14 OFDMA symbols) in the time domain. The smallest resource granularity is represented by the Resource Element (RE) unit that considers 15kHz in frequency and one symbol duration in the time domain. Even if the 3GPP specifications refer to time slots of 0.5ms, the scheduling process fills the pool of REs with 180KHz at the resolution of 1ms. When Multiple-In-Multiple-Out (MIMO) techniques are used, the LTE radio interface combines the facilities of MIMO and OFDMA in what is called the multi-user MIMO technology [36].

The link and adaptation procedure refers to the possibility of adapting the Modulation and the Coding Scheme (MCS) for the transmission based on channel estimations received from each UE. It would be too inefficient if the channel estimation is achieved and reported on 15KHz basis. Therefore, in LTE the channel estimation is achieved based on RB granularity known as a CQI report.

According to [35], two transmission modes are supported in LTE: Time Division Duplex (TDD) and Frequency Division Duplex (FDD). The TDD mode considers a number of 10 consecutive TTIs necessary for the LTE frame. Based on the frame representation, different sub-frames (TTIs) support different uplink-downlink configurations for the whole bandwidth. In contrast, FDD allocates different portions of the frequency spectrum to uplink and downlink transmissions. Because this study concentrates only on the downlink scheduling procedures, the FDD mode is used for the downlink transmission.

2.8 Radio Resource Management in LTE

The RRM entity aims to optimize the problem of allocating the available radio resources in an efficient way, assuring at the same time the satisfaction of end-to-end users according to their QoS requirements. It covers the optimization problems which are not entirely covered by 3GPP specifications in order to be designed by the mobile operators or vendors for their own needs. The network optimization issues are possible on the strength of different adaptation techniques.

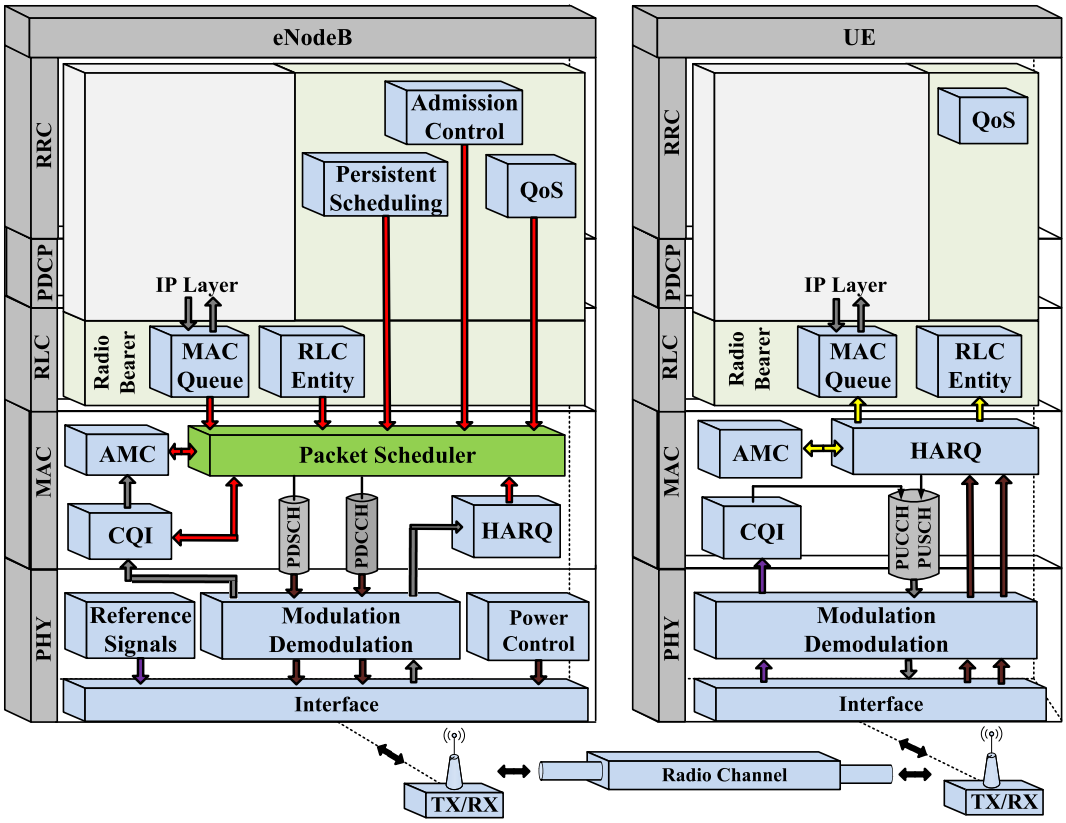


Fig. 2.6 Interaction of the Main RRM Adaptation Techniques

Parts of these techniques were already introduced in the previous sections but are re-stated here in order to highlight the interaction of the LTE packet scheduler with other RRM entities as shown in Fig. 2.6 [20]:

- RRC adaptive techniques: Radio Admission Control (RAC), persistent and semi-persistent scheduling and QoS management;
- RLC adaptive techniques: management of MAC queues, RLC entities;
- MAC adaptive techniques: HARQ processes, AMC entity, dynamic packet scheduler and CQI reports quantization;
- PHY adaptive techniques: Physical Downlink Control CHannel (PDCCH) adaptation, SINR measurements and power control.

It is important to note that the RRC and RLC RRM adaptive functionalities serve the radio bearer requirement. For different application requirements per one UE entity, multiple RRC-RLC RRM functions are considered by the dynamic packet scheduler. Alongside of all of these RRM entities, the dynamic packet scheduler plays a crucial role in the RRM optimization problems. It takes the whole responsibility for selecting different radio bearers to be scheduled in order

to increase the total system capacity and to respect the QoS requirements. The scheduling decision is based on the interaction with other RRM entities. For this reason, the RRM optimization problem is often called as a *cross layer optimization technique*. In other words, the cross layer technique aims to optimize the usage of radio resources for different types of applications, acting as a *bridge* between PHY, MAC, RLC and RRC layers. The major blocks involved in the LTE scheduling decisions are described below:

- **RAC procedure** decides if a new radio bearer is accepted or not in the scheduling procedure. The feasibility of the LTE scheduler can be affected when accepting or rejecting some radio bearers.
- **Semi-persistent scheduling** is used to avoid large control overheads associated with the small data packets such as VoIP [24], [29]. In contradistinction to dynamic scheduling, the semi-persistent scheduling aims to allocate persistent or specific radio resources for the voice services regardless the CQI conditions.
- **CQI Reports:** The eNodeB station generates periodically the Sounding Reference Signals (SRS) which are well known for each active UE. At the PHY layer, each user measures the level of SINR based on the received power of SRS signals. The UE MAC layer applies the quantization process obtaining 4 bit CQI value and then, periodically, UE sends the CQI value to the eNodeB. It is important to note that the CQI value is determined per RB basis (full-band CQI reporting scheme) or for the whole bandwidth (wideband CQI reporting scheme).
- **PDCCH and PDSCH Channels:** Each UE receives the downlink information by using the Physical Downlink Shared Channel (PDSCH). As the name suggest, PDSCH is shared among the active users in the cell. The PDCCH transmission is permitted to consume only part of the used spectrum and for a given number of OFDM symbols in the radio resource grid [20]. The downlink control information in LTE is transmitted by using three control channels, but the most important one for this study is the PDCCH channel. The PDCCH channel carries the user assignments for each RB and the MCS scheme. Another important message transported in

the PDCCH channel is the Downlink Control Information (DCI) that contains different information about the system configuration [20]. The PDCCH adaptation refers to the possibility of reducing the control overhead by using the transmission of DCI formats with lower rates. For the uplink direction, two physical channels are involved in the downlink scheduling procedure: Physical Uplink Control CHannel (PUCCH) and Physical Uplink Shared CHannel (PUSCH). The PUSCH channel is used to transmit the payload in the uplink transmission when an UE is selected by the packet scheduler.

- **Adaptive Modulation and Coding:** After performing the scheduling decision, the transmission parameters for the allocated RBs are required. The AMC provides the MCS schemes of the allocated RBs to the scheduler and then these schemes are transmitted through PDCCH channel. At the beginning of each TTI, the scheduler has to decide what kind of information it should schedule: transmissions or HARQ retransmissions. The primary role of AMC module is to determine, based on the CQI reports, the most suitable MCS scheme for the allocated RBs in order to increase the cell spectral efficiency under a given requirement of the Block Error Rate (BLER). This entity is entitled Inner Loop AMC (ILAMC). This way, users which are located at the cell edge receive lower bit rates whereas users being located near eNodeB experience much higher bit rates. The second role of AMC module adapts the BLER requirements for the first transmission based on the previous HARQ acknowledgements as indicated in [37]. In fact, the retransmissions can support tolerable BLER as the receiver is able to decode the correct version of the transmitted packet based on the combination of different received versions in the past [38]. The entity is entitled Outer Loop AMC (OLAMC).
- **Adaptive Power Control:** In LTE, AMC and Adaptive Power Control (APC) can be used together in order to compensate the unfavorable variations of SINR levels as indicated by the specifications of PHY layer procedures [39]. For example, the downlink power level can be increased for the scheduled users being located at the cell edge in order to improve

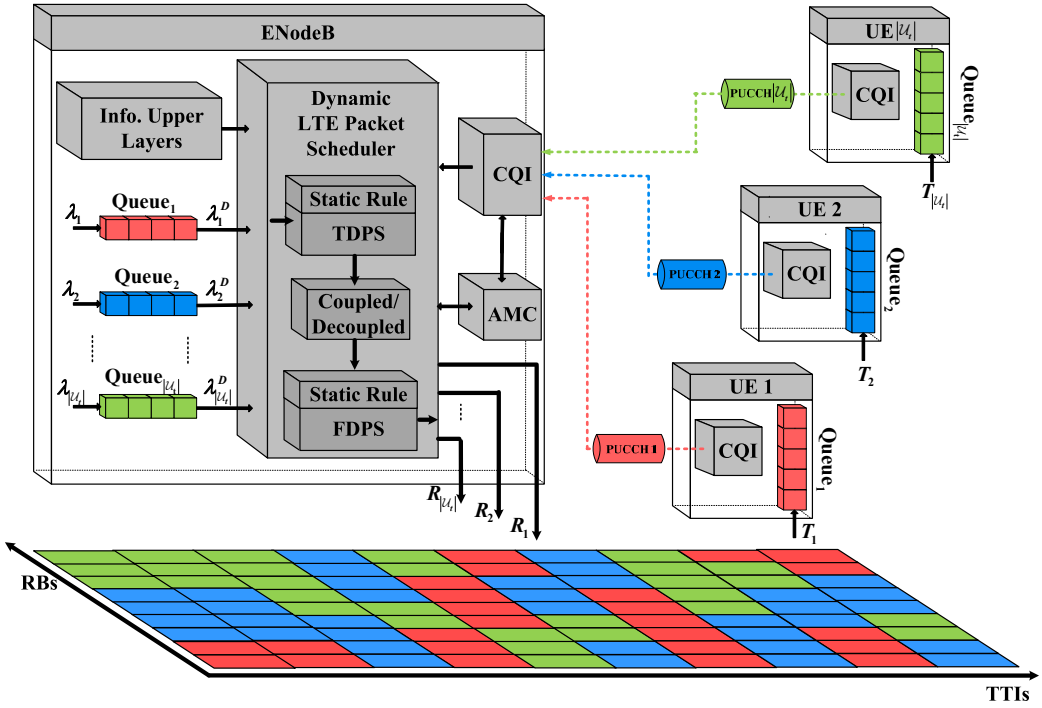


Fig. 2.7 Basic Concepts of the Downlink LTE Packet Scheduler

the supported MCS schemes and implicitly its channel quality, and it can be reduced for scheduled users which are located near eNodeB without decreasing the MCS scheme in order to save the downlink transmission energy. For this research, it is considered that the transmission power is constant for the entire bandwidth and the scheduler performs only the AMC procedures.

2.9 The Dynamic LTE Packet Scheduler

The general overview of the LTE downlink scheduler is depicted in Fig. 2.7. The main responsibility of the LTE downlink dynamic packet scheduler is to allocate a number of pre-selected flows in the time-frequency domain in order to satisfy the multi-objective criterion of the optimization problem. The multi-objective optimization problems will be analyzed in Chapter 3 and the LTE scheduler model is presented in the following sub-section.

Let us consider $|\mathcal{U}_t|$ the number of active UEs in the cell, where $\mathcal{U}_t$ is the set of active users at TTI t and N_F the number of radio bearers or the associated

data flows for each user $i \in \mathcal{U}_t$, where $N_F \geq |\mathcal{U}_t|$. For simplicity, Figure 2.7 considers the special case when $N_F = |\mathcal{U}_t|$. At the RLC level, each transmission MAC queue considers different instantaneous arrival rates $\lambda_i[t_\lambda]$ for each data flow or user $i \in \mathcal{U}_t$, where t_λ represents the time instant when the data packet is arriving in the MAC queue. The scheduler decision influences the size of each MAC queue based on which user $i \in \mathcal{U}_t$ is decided to be scheduled at TTI t, $\forall t = 1,...,N_{TTI}$ and $t_\lambda \leq t$, where N_{TTI} is the total number of TTIs being considered for a given transmission session. The instantaneous departure rate $\lambda_i^D[t]$ for user $i \in \mathcal{U}_t$ represents the total number of bits that can be transmitted from queue *i* at TTI t if user $i \in \mathcal{U}_t$ is selected for scheduling. The instantaneous departure rate has a close connection with the CQI reports and the AMC module since the total number of transmitted bits is directly connected with the supportable number of bits that can be transmitted on different RBs.

Let us use $|\mathcal{B}|$ to denote the total number of RBs in a given LTE bandwidth, where $\mathcal{B}$ is the set of RBs. Based on the scheduling rule, different users are selected to transmit on different RBs. The instantaneous achievable rate $r_{i,j}[t]$ represents the total number of bits that can be transmitted in the downlink direction for user $i \in \mathcal{U}_t$ and for each RB $j \in \mathcal{B}$ at TTI t before performing the scheduling decision. If user $i \in \mathcal{U}_t$ is selected to transmit on multiple RBs, the obtained rate is $R_i[t] = \sum_{j \in \mathcal{B}} r_{i,j}[t], \forall i \in \mathcal{U}_t$, being entitled the instantaneous user rate. At the reception side, if the transmitted bits in the previous TTI were successfully decoded (HARQ/RLC acknowledgement), then the instantaneous rate for the scheduled user $i \in \mathcal{U}_t$ becomes the user throughput $T_i[t]$.

In LTE, for each RB $j \in \mathcal{B}$, the user $i \in \mathcal{U}_t$ which maximizes a specific scheduling metric based on specific scheduling rule is selected to transmit on that RB. It can be observed that the scheduler acts as an interface and maps packets from the MAC queue into the time-frequency resource grid based on specific scheduling metrics. The RB allocation can take different forms by exploiting both

time and frequency domains, leading to different operation modes of the LTE scheduler as shown in Fig. 2.7 such as coupled or decoupled time-frequency scheduling.

2.9.1 Operation Modes in LTE Packet Scheduling

The LTE packet scheduling process is governed by two main concepts: *the scheduling rule* and the *scheduling procedure*. As mentioned in Chapter 1, the scheduling rule can take two main forms:

- **Static Scheduling Rule (SSR)**: The same scheduling rule is applied at each TTI *t* for the entire transmission;
- **Dynamic Scheduling Rule (DSR)**: Different scheduling rules may be applied at each TTI forming a policy of scheduling disciplines. In fact, the dynamic behavior of different scheduling rules is the scope of this study and the techniques behind of this approach are analyzed in Section 2.10.

The scheduling procedure considers two stages: *user selection* and *resource blocks allocation*. These stages have to be processed in less than 1*ms,* and then, the scheduling procedure can work under the following three different modes at each TTI:

- **Active Selection of UEs and Passive Allocation of RBs**: The user selection is achieved at each TTI and the winner takes the whole bandwidth. It is considered in this mode that the RB allocation stage is suppressed. For these reasons, the packet scheduling process is entitled Time Domain Packet Scheduling (TDPS) due to the fact that different users are scheduled at different TTIs without any consideration of the frequency domain.
- **Passive Selection of UEs and Active Allocation of RBs**: The user selection is performed just once at the beginning of the transmission. Therefore the user selection stage is suppressed. The allocation of RBs is achieved at each TTI, and the radio resources are allocated only for the set of preselected users. Under these circumstances, the LTE scheduling is

considered to be Frequency Domain Packet Scheduling (FDPS) and aims to allocate in frequency at each TTI all users which are preselected at the beginning of the transmission session.

- **Active Selection of UEs and Active Allocation of RBs:** At the beginning of each TTI, a group of UEs are selected to transmit on different RBs. The packet scheduling process considers the facilities of both TDPS and FDPS stages and it can be entitled the TDPS/FDPS packet scheduling.

The way how the TDPS/FDPS packet scheduling procedure is computed, divides the scheduling procedure into two main categories [20]:

- **Coupled TDPS/FDPS**: The selection of UEs and the allocation of RBs are computed at the same time at each TTI.
- **Decoupled TDPS/FDPS**: The selection of UEs is performed first at each TTI, and based on the selected UEs, the allocation of RBs is performed.

By using different scheduling procedures, one scheduling rule is considered for TDPS, FDPS and coupled TDPS/FDPS scheduling whereas two disciplines are needed for the decoupled TDPS/FDPS scheduler (one scheduling rule for the selection of UEs and one rule for RB metric calculations) [20], [40]. Based on the principles exposed above, the LTE packet scheduler can performs under different modes depending on the dynamicity of scheduling rules and procedures types:

***Existing LTE Scheduling Modes* [20],[40]:**

- **TDPS-SSR** – A static rule is performed and only one user is selected at each TTI based on the wideband system CQI report.
- **FDPS-SSR** – The static scheduling rule is applied at each TTI in the frequency domain in order to schedule the same group of users.
- **Coupled TDPS/FDPS-SSR** – To exploit the OFDMA advantages, a static scheduling rule is applied and users are allocated in both time and frequency domains.
- **Decoupled TDPS-SSR/FDPS-SSR** – Users are allocated in both time and frequency domains. Two static scheduling rules are applied (different rules for TDPS and FDPS) and the scheduling objectives are shared between TDPS and FDPS domains [40].

Proposed LTE Scheduling Modes:

- **TDPS-DSR** – Different scheduling rules are applied at each TTI and only one user with the best scheduling metric is selected to transmit at different TTIs for the entire frequency domain.
- **FDPS-DSR** – Same as FDPS-SSR but with different rules being applied at each TTI in the frequency domain;
- **Coupled TDPS/FDPS-DSR** – Users are scheduled in both time and frequency domains under with a dynamic scheduling rule.
- **Decoupled TDPS-DSR/FDPS-SSR** – The user selection stage is governed by a dynamic rule whereas the allocation of RBs is achieved by using a static discipline.
- **Decoupled TDPS-SSR/FDPS-DSR** – A static rule is used for the prioritization of UEs and a dynamic discipline for the FDPS domain.
- **Decoupled TDPS-DSR/FDPS-DSR** – The scheduling rules take the dynamic form for both TDPS and FDPS domains.

It is important to point out that in the case of decoupled TDPS/FDPS modes, an intermediary step between TDPS and FDPS scheduling is necessary in order to validate if there is enough PDCCH resources for transmission for the pre-selected users at each TTI. However, in this research, the coupled TDPS/FDPS-DSR principle is studied. The key concepts of the proposed architecture are presented in Section 2.10 and in the following sub-section are discussed the main characteristics of the coupled TDPS/FDPS-DSR scheduling principle.

2.9.2 Coupled TDPS/FDPS-DSR Scheduling

The coupled TDPS/FDPS scheduling introduces the dynamicity in both time and frequency domains. For this reason, this scheme is entitled *dynamic packet scheduling technique*. When the coupled TDPS/FDPS-DSR is performed, the dynamicity is introduced also in the selection procedure of scheduling rules. The set of selected users based on the metric prioritization is denoted by $\mathcal{U}_t^{TF} \subseteq \mathcal{U}_t$. Then, the mathematical representation of the coupled TDPS/FDPS-DSR principle

$$\begin{cases} \bigcup_{i \in \mathcal{U}_t^{TF}} \mathcal{B}_{i,t} = \mathcal{B} \\ \mathcal{B}_{i,t} \subseteq \mathcal{B}, \forall i \in \mathcal{U}_t^{TF} \\ \mathcal{B}_{i_1,t} \bigcap \mathcal{B}_{i_2,t} = \varnothing, \forall i_1 \in \mathcal{U}_t^{TF}, \forall i_2 \in \mathcal{U}_t^{TF} \\ \mathcal{U}_t^{TF} \subseteq \mathcal{U}_t \end{cases} \quad (2.1)$$

at each TTI t is denoted by Eq. 2.1, where $\mathcal{B}_{i,t}$ represents the list of RBs which should be allocated to UE $i \in \mathcal{U}_t$. The coupled or joint TDPS/FDPS represents a very powerful improvement due to the OFDMA technique which permits to increase the total cell spectral efficiency. The key factor of joint TDPS/FDPS is to exploit the variations into the SINR levels due to the interference with other cells, fast-fading, path and penetration losses. In this sense, users with deep fades are avoided, and then only those UEs that experience very good frequency selective channel qualities are possibly scheduled. Recent studies show that the multi-user diversity gain with coupled TDPS/FDPS scheduling is able to achieve 40% gain from the system capacity point of view when compared with TDPS for best effort users, Poisson arrival rates and 1x2(1 transmitter,2 receivers) MIMO scheme [20]. When the number of active users is large, the computational complexity $\left(|\mathcal{U}_t| \times |\mathcal{B}|\right)$ becomes a concern in the coupled TDPS/FDPS – DSR scheme.

2.10 The Integration of the Proposed Scheduling Architecture in the RRM Environment

For the downlink scheduling purpose, the eNodeB station receives data packets from the IP layer (Fig. 2.8) and computes the MAC queues for each active user (and for each active data flow). The RLC layer performs the segmentation and the concatenation procedures in order to calculate the Transport Block Size (TBS) for each scheduled user. The TBS represents the number of bits which can be sent to the scheduled users based on the supportable MCS schemes. More details about the TBS computation in LTE downlink scheduling are provided in Appendix B. The Transport Block (TB) format is associated with the HARQ module which can decide to retransmit the entire TB in the case of the erroneous

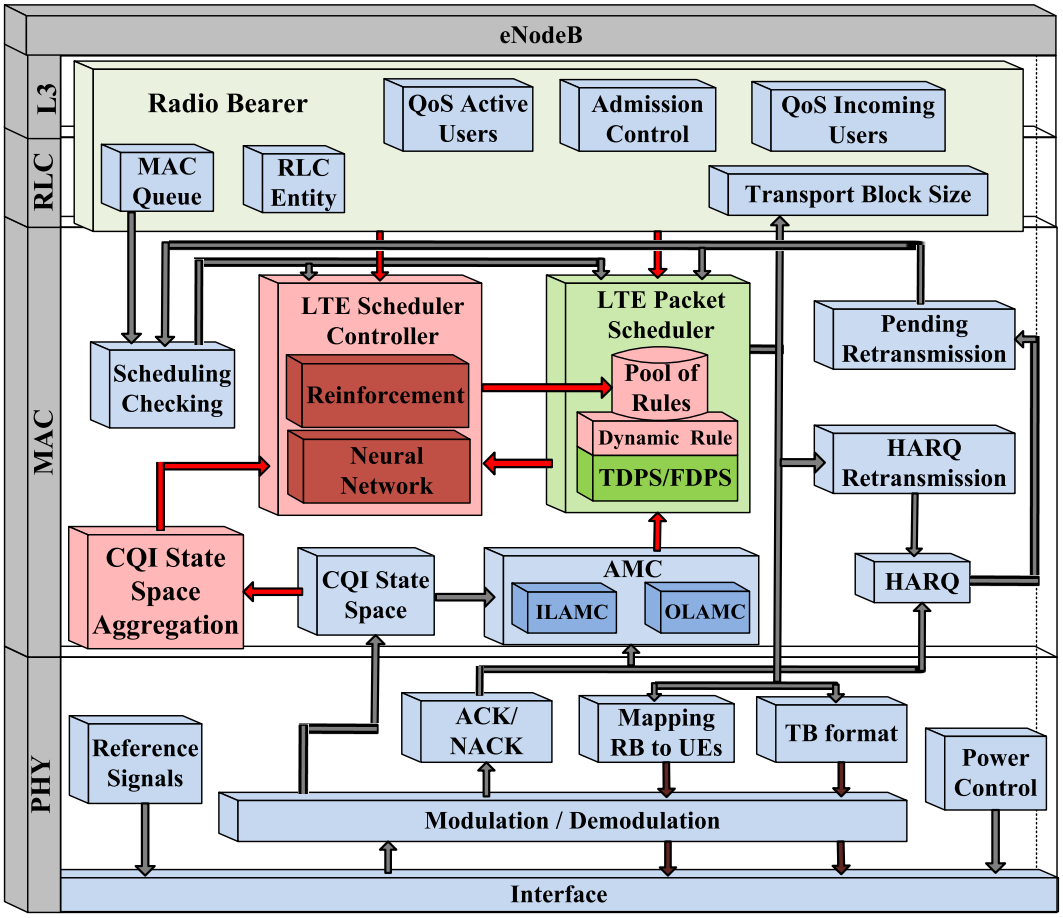


Fig. 2.8 The Proposed RRM Architecture

reception. After performing the scheduling procedure, the supportable MCS is assigned for each scheduled user and for each RB. The PDCCH channel contains the assigned RBs to different scheduled users and the formats of TBs with the MCS schemes and then, the modulation procedure is performed. Basically, at the reception side, the scheduled users access the PDSCH channel based on the information received on the PDCCH physical channel.

In the case of the erroneous reception (HARQ NACK over PUCCH), the packet scheduler has to schedule new transmissions and pending HARQ retransmissions. The mix of retransmissions and new transmissions are not permitted for the same user [20]. The PS entity can use different priorities when scheduling the retransmissions. For instance, in the FDPS domain, the best RBs are allocated for the new transmissions whereas the rest of RBs are allocated for the retransmissions [20]. In this case, the Scheduling Checking block validates if there is enough PDCCH resources for the scheduled users.

The CQI state space module (Fig. 2.8) collects the CQI reports from each active user. Based on this entity and based on the ACK/NACK block, the AMC module can adapt the MCS scheme and the BLER target in order to assign supportable schemes for the assigned RBs. More details about the CQI cycle in LTE networks and AMC techniques are provided in Appendix B.

As mentioned earlier, the RAC entity can accept (incoming QoS constraints) different flows to be scheduled if the DSR-SMOO/CMOO problems permit. Otherwise, the RAC entity can reject (existing QoS constraints) different flows in order to reach the scheduler optimality. The RAC procedure represents a key factor in assuring the optimality for different DSR-SMOO/CMOO problems.

The DSR-SMOO/CMOO problems constitute difficult tasks from the coupled TDPS/FDPS-DSR implementation point of view. The main difficulty is to find the most suitable scheduling rule at each TTI. The available information which can help in selecting proper scheduling discipline is the scheduler state space in terms of different parameters such as user data rates, packet delays, arrival rates, packet loss rates and CQI reports. More details about the scheduler state space are given in Chapter 3.

The integration of the proposed coupled TDPS/FDPS-DSR scheduling technique in the RRM architecture is represented in Fig. 2.8 in which an intelligent controller is needed in order to refine the set of sustainable scheduling policies for the DSR-SMOO/CMOO problems. The mathematical optimization for the coupled TDPS/FDPS-DSR architecture is proposed in Chapter 3.

The best way to find the most suitable scheduling rule for a given state is to interact with the RRM environment. The behaviors of the LTE scheduler and RRM entities are totally unknown when different scheduling rules are applied. In order to learn the behavior of RRM entities under various scheduling rules, the RL methodology is used in this study. The RL approach has been developed in many controlling problems to provide promising results [42]. The basics of RL approach imply the interaction between an intelligent agent called LTE controller and the unknown environment entitled the LTE packet scheduler (including here other RRM functionalities). The interaction procedure is modeled based on the

scheduler state space which is provided to the LTE controller at each TTI t from the RRM environment. Based on the input state, the LTE controller selects, according to the policy learned so far, the action or the scheduling rule in order to select the set of users $\mathcal{U}_t^{TF} \subseteq \mathcal{U}_t$ to be scheduled in the frequency domain. As a response, the RRM environment evaluates the scheduling performance of the applied action at TTI t and provides the reward value to the LTE controller at TTI t+1. The reward value is a measure of performing a given scheduling rule based on different scheduler states. The reasoning behind of this approach is to collect as many rewards as possible for each state and to select the action that maximizes the accumulated reward value in order to increase the number of optimal/feasible states for a given DSR-SMOO/CMOO problem. More precisely, the controller role is to form a sustainable set of policies which consist of different actions that can maximize the accumulated reward for each given input controller state. The first proposals of applying the RL concepts in LTE scheduling can be found in [43], [44]. To conclude, the RL methodology is used in LTE scheduling in order to solve DSR-SMOO/CMOO problems. More details about the integration of the RL concept in the considered optimization problems are presented in Chapter 3.

Another major concern in the proposed approach is the scheduler state space dimensionality. Due to the very large dimension of the input scheduler state, the exploration stage requires more time to sweep the entire scheduler state space. Then, the aggregation procedure is needed in order to reduce the size and the dimension of this space. In Fig. 2.8, an aggregation block is proposed for the CQI state space compaction based on pre-processing, classification and regression procedures. For other input scheduler parameters, some statistical models can be applied in order to extract the relevant features. Chapter 4 proposes innovative concepts of aggregating the entire scheduler state space.

Moreover, the scheduler state space keeps continuous after the aggregation procedure, which implies in fact the impossibility of exploring the entire aggregate scheduler states. This way, the RL algorithm with the approximation function is proposed as a novel technique to approximate the state and state-action values. The MLPNN approach is used in this study to approximate a proper

scheduling rule for each scheduler state at each TTI. The MLPNN functions are trained based on the interaction with the reinforcement block, which in fact determines the type of RL algorithm. The RL approach is based on two stages: *exploration* and *exploitation*. In the exploration stage, the MLPNN functions are trained based on the reinforcement value. At this stage, the scheduling policy is refined. The exploitation stage analyses the performance of the trained MLPNN functions when applying the obtained set of scheduling policies. The idea is to obtain sustainable scheduling policies which can maximize in the long term purpose the number of feasible TTIs under various conditions. Precise details about the controller architecture for different RL algorithms under continuous state spaces are provided in Chapter 5. The set of sustainable scheduling policies for different DSR-SMOO/CMOO problems are obtained based on the architecture exposed in Fig. 2.8 and their performances are analyzed in Chapters 6 and 7.

2.11 Summary

The role of the LTE scheduler is to reach the optimal or feasible state that can guarantee the total cell spectral efficiency maximization by respecting, at the same time, the user fairness criterion and the QoS requirements, and keeping the data queues stable. The study on how the scheduler is able to reach the optimal state and to keep the system as long as possible in the optimality region has attracted a big interest. By using different scheduling rules being oriented on different objectives based on the coupled TDPS/FDPS-DSR architecture, the number of TTIs, when the scheduler is declared feasible, can be increased when compared with formal architectures. The proposed architecture makes use of an intelligent controller which is able to interact with the LTE scheduler and RRM entities. Based on the received state and the reward value, a proper scheduling rule from a given pool of disciplines is selected in order to make the system stay in the optimum state. The reinforcement learning with the neural network as a function approximation is used in order to produce sustainable and optimal scheduling policies which are able to reach the desired state as fast as possible starting from any given initial scheduler state.

Chapter 3

LTE Scheduling Multi-Objective Optimization

3.1 Chapter Outline

The most important elements involved in LTE scheduling multi-objective optimization are the utility and objective functions, and the scheduler state space. The scheduler state space represents the scheduling parameters necessary to compute the utility and objective functions. Utility functions quantify the benefit of allocating radio resources for a particular MOO problem based on a given scheduler state. Different utility functions impact differently in the optimization problem, addressing a particular objective function. The objective function evaluates the scheduling procedure performance based on the selected utility function. The scheduling procedure aims to maximize the sum of user utilities in the long term purpose. In this sense, the scheduling rule which is a form of Marginal Utility Function (MUF) (derivative of utility function) is applied TTI-by-TTI. Based on the type of scheduling rule, the scheduler performance is balanced in the direction of the addressed objective, degrading the scheduler performance from the perspective of other objectives. In order to straighten the balance between objectives, this chapter addresses the aggregate utility function

optimization problem which can be performed by applying different scheduling rules at each TTI (DSR-SMOO/CMOO techniques). Three types of linearization techniques are proposed in this sense. The Augmented Lagrangian principle is used in order to introduce the set of objective constraints in the aggregate optimization problem. Due to the high complexity overhead involved in the optimal optimization problems, a sub-optimal scheduler is proposed by dividing the entire non-linear problem in two linear optimization problems. The first optimization problem aims to select the best scheduling rule which can maximize the multi-objective performance. In the second optimization problem, the radio resource assignment is performed based on the selected scheduling rule. The RL based LTE scheduling is proposed to learn and to refine the policies of scheduling rules in order to reach optimal or near-optimal solutions of the proposed aggregate MOO problem. Based on the proposed DSR-SMOO/CMOO principles, a novel classification of scheduling techniques is proposed. The most relevant related work is analyzed in order to highlight the necessity of the proposed approach.

3.2 LTE Scheduling Process Components

The scheduler process considers four main components: scheduling procedure, scheduler state space, MOO performance evaluation and the scheduling rule. The scheduling procedure includes the user selection, resource allocation and MCS assignment stages governed by the scheduling discipline. The scheduler state space represents a collection of parameters and indices that can be used for the scheduling procedure and for the multi-objective performance evaluation (e.g. arrival rates or instantaneous user rate, as depicted in Fig 2.7).

The most pretentious task in OFDMA scheduling is to find the potential benefit of using certain radio resources for each UE $i \in \mathcal{U}_t$ and for a given performance criterion. More precisely, being given a certain performance criterion, the scheduler should be aware about the exact price or cost value of allocating RB $j \in \mathcal{B}$ to UE $i \in \mathcal{U}_t$ for its target objective satisfaction. So, the scheduler is responsible for optimizing the obtained pricing structure problem.

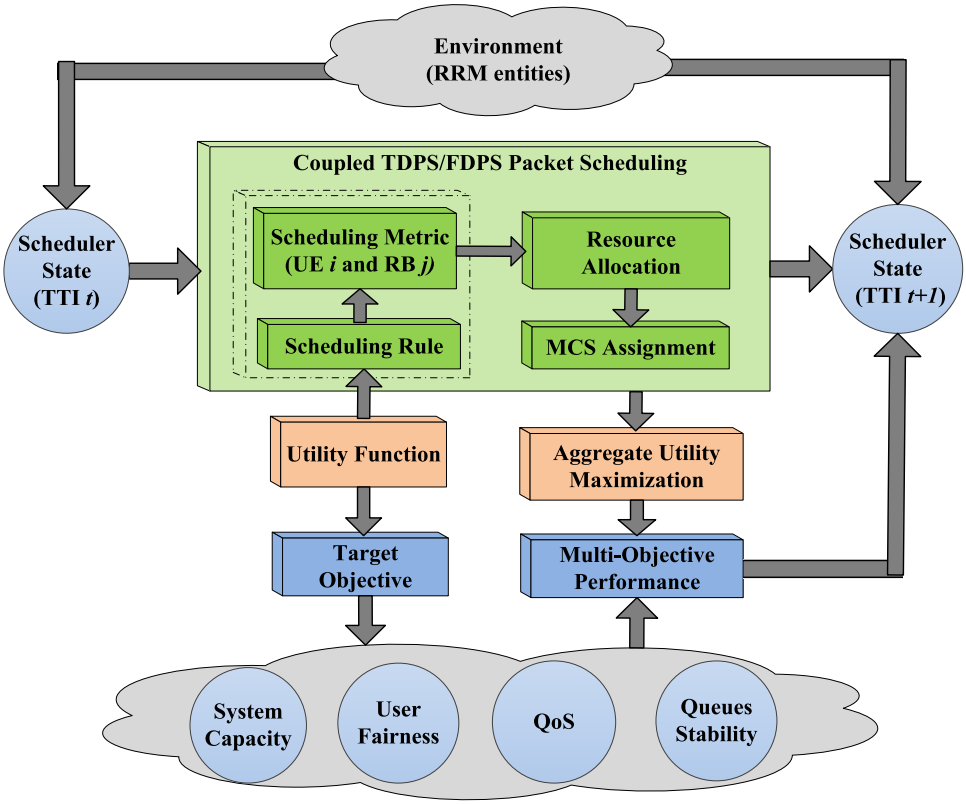


Fig. 3.1 The Interface Between the Coupled TDPS/FDPS-SSR Packet Scheduler and the Multi-Objective Optimization Problem

The potential benefit quantification of using some limited resources is inherited from the utility theory in economics which has been applied with great success in wireless networks in order to guarantee the QoS requirements and to exploit the multi-user diversity principle in opportunistic scheduling [45]. In LTE networks, the proposed scheduling procedure maps the performance criteria in some utility metrics for each user $i \in \mathcal{U}_t$ and for each RB $j \in \mathcal{B}$. Then, the instantaneous optimization problems resume to the sum maximization of each user utility TTI-by-TTI.

Adopting the performance criteria in order to evaluate the performance of user centric objectives for different type of services represents a crucial task. As mentioned earlier, by using classical scheduling procedures, it is difficult to reach the optimality of multiple objectives simultaneously. Therefore, some priorities in satisfying particular objectives are given by adopting different performance criteria at once. However, the performance criteria denote in fact the types of utility functions. For instance, if the utility function addresses the HoL packet

delay performance, the scheduler is designed in such a way that the packet delay budget should be satisfied in certain requirements. If the optimality of the first condition is satisfied, then other objectives can be considered depending on the particularity of the utility function. Based on the type of the objective, the coupled TDPS/FDPS-SSR scheduler determines the price value of allocating RB $j \in \mathcal{B}$ to user $i \in \mathcal{U}_t$ (Fig. 3.1). Then, the scheduling decision is performed, and the impact of its decision into the multi-optimization problem can be evaluated through the multi-objective performance block. The optimality condition is reached if the sum of the allocated user utilities is maximized under different objective constraints. When performing the scheduling decision at TTI t, the scheduler is able to evolve to the newest state at TTI t+1. The details of the objective and utility functions will be given in Section 3.5 after the introduction on the scheduler state space.

3.3 The LTE Packet Scheduler State Space

An important role in LTE scheduling is played by the scheduler state space since both optimization approaches, SMOO and CMOO, perform the scheduling procedure at each TTI based on the scheduler conditions. Without going through precise details at this stage, the scheduler state space is exploited in different ways based on the exploited optimization type:

1. In the SSR-SMOO problems, the scheduler state space provides the necessary parameters for the utility function computation;
2. For the DSR-SMOO/CMOO problems, alongside the provision of utility parameters, proper scheduling rules are selected TTI-by-TTI based on the scheduler state in order to develop sustainable scheduling policies.

Inevitably, the selected scheduling rule affects part of the scheduler state space evolution. The scheduler state space is divided into two disjoint subspaces:

- ***Uncontrollable scheduler state space parameters***: The CQI reports, HARQ indicators, arrival bit rates and QoS requirements are included.
- ***Controllable scheduler state space parameters***: The parameters which are responsible for the objective performance evaluation, such as HoL packet

delay, average user rate, normalized user rate, packet loss rate and queue size are included. More details about these indicators will be provided gradually in the following sub-sections.

The objective performance indicator is included in the scheduler state space, and for these reasons the multi-objective optimality is directly regarded to the scheduler state optimality. Let us define the scheduler state $\mathcal{S}_t^S \in \mathbb{R}^{D\left[\mathcal{S}_t^S\right]}$ at TTI *t*, where $D\left[\mathcal{S}_t^S\right]$ denotes the scheduler state space dimension which depends on the number of active users from $\mathcal{U}_t$. The scheduler state $\mathcal{S}_t^S$ is divided into $\mathcal{N}_S$ number of disjoint subsets based on different performance parameters achieved by each UE $i \in \mathcal{U}_t$. Then, the state space dimension can be defined based on Eq. 3.1:

$$D\left[\mathcal{S}_t^S\right] = \mathcal{N}_S \cdot \left|\mathcal{U}_t\right| \tag{3.1}$$

Based on the impact of the scheduling discipline, the scheduler state space representation is obtained based on the reunion of two disjoint subspaces as expressed by Eq. 3.2:

$$\mathcal{S}_t^S = \mathcal{S}_t^{S,C} \bigcup \mathcal{S}_t^{S,U} \tag{3.2}$$

where $\mathcal{S}_t^{S,C}$ is the controllable scheduler state space, whereas $\mathcal{S}_t^{S,U}$ is the uncontrollable scheduler state space. When the scheduling procedure is performed, the scheduler evolves from $\mathcal{S}_t^S$ to $\mathcal{S}_{t+1}^S$. The $\mathcal{S}_{t+1}^{S,U}$ subspace is the result of stochastic processes rather than the results of the previous scheduling procedure being applied in state $\mathcal{S}_t^S$. The components of the scheduler state space $\mathcal{S}_t^S$ are illustrated and analyzed in the following sub-sections.

3.3.1 The Uncontrollable Scheduler State Space

The uncontrollable scheduler state space comprises indices and parameters which reflect mainly the channel conditions, the service parameters from the upper layers, and the QoS requirements for each active data flow. Even if these parameters are modeled as random processes rather than the scheduling procedure

results, the obtained subspace plays a crucial role in achieving and maintaining the feasible regions of different objectives (more details in Chapters 6 and 7). The uncontrollable state space $\mathcal{S}_t^{S,U}$ encompasses the following elements:

1. **Channel Quality Indicator (CQI) Reports**: It is assumed that at each TTI *t,* each UE $i \in \mathcal{U}_t$ reports, without any delay, the CQI value for each RB $j \in \mathcal{B}$ through the PUCCH control channel. The transmission on PUCCH is considered to be errorless. Let us define $CQI_{i,j}[t]$ as the CQI report value for the resource $j \in \mathcal{B}$ and user $i \in \mathcal{U}_t$ at TTI *t* and $\mathcal{S}_{i,t}^{CQI} = \bigcup_{j=1}^{|\mathcal{B}|} CQI_{i,j}[t]$ is the CQI vector for UE $i \in \mathcal{U}_t$. Then the overall CQI set for all active users can be defined as follows:

$$\mathcal{S}_{1,t}^{S,U} = \mathcal{S}_t^{CQI} = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{S}_{i,t}^{CQI} \tag{3.3}$$

2. **Achievable instantaneous user rate**: Based on the $CQI_{i,j}[t]$ report, the instantaneous achievable user rate $r_{i,j}[t]$ is computed for the scheduling decision. Following the same principle, the achievable rate set for UE $i \in \mathcal{U}_t$ at TTI *t* is defined as $\mathcal{S}_{i,t}^{r} = \bigcup_{j=1}^{|\mathcal{B}|} r_{i,j}[t]$. The fully observable achievable user rate subset for the LTE scheduler takes the following form:

$$\mathcal{S}_{2,t}^{S,U} = \mathcal{S}_t^{r} = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{S}_{i,t}^{r} \tag{3.4}$$

3. **HARQ notifications**: These refer to a binary value $HARQ_{i,t} = \{0,1\}$ which decides if eNodeB has to re-transmit packets to the scheduled UE in the previous TTI (1 -re-transmission, 0 -no re-transmission). The HARQ decision state for each UE is denoted by Eq. 3.5.

$$\mathcal{S}_{3,t}^{S,U} = \mathcal{S}_t^{HARQ} = \{HARQ_i[t]\}, i = 1,...,|\mathcal{U}_t| \tag{3.5}$$

4. **Instantaneous arrival rate:** The set of arrival rates at TTI *t* can be expressed as follows:

$$\mathcal{S}_{4,t}^{S,U} = \mathcal{S}_t^{\lambda} = \{\lambda_i[t] \in \mathbb{R}_+\}, i = 1,...,|\mathcal{U}_t| \tag{3.6}$$

5. **Average arrival bit rate:** By using the instantaneous arrival bit rate $\lambda_i[t]$, the recursive representation is indicated by Eq. 3.7 and the corresponding set is expressed by Eq. 3.8:

$$\overline{\lambda}_i[t] = \left(1-\beta_{\overline{\lambda}}\right)\cdot\overline{\lambda}_i[t-1] + \beta_{\overline{\lambda}}\cdot\lambda_i[t] \tag{3.7}$$

$$\mathcal{S}_{5,t}^{S,U} = \mathcal{S}_t^{\overline{\lambda}} = \left\{\overline{\lambda}_i[t] \in \mathbb{R}_+\right\}, i = 1,...,|\mathcal{U}_t| \tag{3.8}$$

In Equation 3.7, the parameter $\beta_{\overline{\lambda}}$ sets the time window necessary for averaging the instantaneous arrival bit rate.

6. **QoS Requirements**: For each traffic type, a set of QoS requirements is imposed by 3GPP as indicated in Section 2.5. Let us define for each UE *i* from $\mathcal{U}_t$ the QoS requirements as follows:

$$\mathcal{S}_{QoS,i}^{S,U}[t] = \left\{P_i[t], \underline{T}_i[t], \overline{T}_i[t], d_i^{\overline{HoL}}[t], R_i^{\overline{PL}}[t]\right\} \tag{3.9}$$

where $P_i[t]$ is the priority level corresponding to flow (UE) $i \in \mathcal{U}_t$, $\underline{T}_i[t]$ and $\overline{T}_i[t]$ are the minimum and maximum acceptable limits of the user throughput $d_i^{\overline{HoL}}[t]$ is the upper bound of HoL packet delay and $R_i^{\overline{PL}}[t]$ is the maximum acceptable limit for the packet loss rate for a given BLER. It is important to notice that all the QoS parameters are time dependent since UE $i \in \mathcal{U}_t$ can switch from one service to another during the simulation time. The entire set of QoS requirements are denoted by:

$$\mathcal{S}_{6,t}^{S,U} = \mathcal{S}_{QoS,t}^{S} = \bigcup_{i=1}^{|\mathcal{U}_t|}\mathcal{S}_{QoS,i}^{S}[t] \tag{3.10}$$

For a comprehensive representation, the uncontrollable scheduler state space can be defined by unifying the subspaces defined above:

$$\mathcal{S}_t^{S,U} = \bigcup_{p_i=1}^{6}\mathcal{S}_{p_i,t}^{S,U} \tag{3.11}$$

The uncontrollable subspace requires a special attention, especially on the CQI state space $\mathcal{S}_t^{CQI}$ which can improve or deteriorate the system throughput and

user fairness trade-off control. A special aggregation module is proposed in this sense in Chapter 4 in order to extract the relevant information from $\mathcal{S}_t^{CQI}$ to be used for the DSR-SMOO/CMOO problems. Other parameters from $\mathcal{S}_t^{S,C}$ can be compacted by using statistical models as indicated in Chapter 4.

3.3.2 The Controllable Scheduler State Space

The controllable scheduler subspace denotes the set of indices which are used for the multi-objective performance evaluation. Basically, when $\mathcal{S}_t^{S,C}$ is optimal, the entire scheduler state is considered to be in the optimal region. Under these circumstances, the uncontrollable subspace $\mathcal{S}_t^{S,U}$ provides the necessary information for the MOO problem in order to maintain the system as long as possible in the optimal region. The controllable subspace being considered here comprises the following elements:

1. **Instantaneous user rate**: When performing the scheduling procedure, the instantaneous user rate $R_i[t] \in \mathbb{R}_+$ represents the total number of bits associated to the scheduled user $i \in \mathcal{U}_t$. The instantaneous user rate state space is defined based on the following equation:

$$\mathcal{S}_{1,t}^{S,C} = \mathcal{S}_t^R = \{R_i[t]\}, i = 1,...,|\mathcal{U}_t| \tag{3.12}$$

2. **Instantaneous user throughput**: If the transmitted packets in the previous TTI were correctly decoded by each scheduled user ($HARQ_i[t] = 0$), then the instantaneous user rate becomes the instantaneous user throughput $T_i[t] \in \mathbb{R}_+$, and the associated space is represented by Eq. 3.13:

$$\mathcal{S}_{2,t}^{S,C} = \mathcal{S}_t^T = \{T_i[t]\}, i = 1,...,|\mathcal{U}_t| \tag{3.13}$$

3. **Transmission queue size**: For each active flow, a MAC queue is associated in order to be served by the scheduler entity. It was assumed already for simplicity in Chapter 2 that each active UE has only one active flow or radio bearer $(|\mathcal{U}_t| = N_F)$. In this sense, if $q_i^{TX}[t] \in \mathbb{N}_+$ is the transmission queue size

for bearer or UE *i* at TTI *t*, then the set of queue sizes for each user belonging to $\mathcal{U}_t$ is defined as follows:

$$\mathcal{S}_{3,t}^{S,C} = \mathcal{S}_t^{qTX} = \left\{ q_i^{TX}\left[t\right] \right\}, i = 1,...,\left|\mathcal{U}_t\right| \tag{3.14}$$

4. **Instantaneous HoL packet delay ($d_i^{HoL}\left[t\right]$)**: This element represents the maximum waiting time for a given packet in the MAC queue. The data set which represents the HoL delay $d_i^{HoL}\left[t\right] \in \mathbb{R}_+$ for each UE $i \in \mathcal{U}_t$ is given by:

$$\mathcal{S}_{4,t}^{S,C} = \mathcal{S}_t^{dHoL} = \left\{ d_i^{HoL}\left[t\right] \right\}, i = 1,...,\left|\mathcal{U}_t\right| \tag{3.15}$$

5. **Packet Loss Rate ($R_i^{PL}\left[t\right] \in \mathbb{R}_+$)**: This indicates the number of lost packets in a given time window T_w^{PLR}. The corresponding set of PLRs is denoted by the following equation:

$$\mathcal{S}_{5,t}^{S,C} = \mathcal{S}_t^{PLR} = \left\{ R_i^{PL}\left[t\right] \right\}, i = 1,...,\left|\mathcal{U}_t\right| \tag{3.16}$$

6. **Reception queue size ($q_i^{RX}\left[t\right] \in \mathbb{N}_+$)**: The scheduler may take into consideration the UE buffer status in order to avoid the overflow effect. The reception queue size set is denoted by:

$$\mathcal{S}_{6,t}^{S,C} = \mathcal{S}_t^{qRX} = \left\{ q_i^{RX}\left[t\right] \right\}, i = 1,...,\left|\mathcal{U}_t\right| \tag{3.17}$$

7. **Average user throughput**: It is used to improve the fairness among users. If the instantaneous user throughput $T_i\left[t\right] \in \mathbb{R}_+$ is used as the fairness satisfaction metric, then the scheduler should be fair at each TTI. This aspect is undesirable because it affects the spectral efficiency. Therefore, it is preferred to evaluate the fairness performance by using a time window or a predefined number of TTIs. So, the average user throughput $\overline{T}_i\left[t\right]$ is defined as follows:

$$\overline{T}_i\left[t\right] = \left(1 - \beta_{\overline{T}}\right) \cdot \overline{T}_i\left[t-1\right] + \beta_{\overline{T}} \cdot T_i\left[t\right] \tag{3.18}$$

where $\beta_{\overline{T}}$ represents the forgetting factor which impacts in the scheduler performance. The lower values for parameter $\beta_{\overline{T}}$ implies in fact lower impacts of the current scheduling procedure in the optimization problem. More details about the types of averaging filters and their effects in the DSR-

SMOO/CMOO performance are largely discussed in Chapter 6. The average user throughput set is represented by Eq. 3.19:

$$\mathcal{S}_{7,t}^{S,C} = \mathcal{S}_t^{\overline{T}} = \left\{ \overline{T}_i[t] \right\}, i = 1,...,|\mathcal{U}_t| \tag{3.19}$$

8. **The average transmission queue size** is computed in a similar way to $\overline{T}_i[t]$, and some particular types of scheduling rules consider the online computation of average transmission queue size $\overline{q_i^{TX}}[t] \in \mathbb{R}_+$ such that:

$$\overline{q_i^{TX}}[t] = \left(1 - \beta_{\overline{q^{TX}}}\right) \cdot \overline{q_i^{TX}}[t-1] + \beta_{\overline{q^{TX}}} \cdot q_i^{TX}[t] \tag{3.20}$$

with the corresponding subset:

$$\mathcal{S}_{8,t}^{S,C} = \mathcal{S}_t^{\overline{qTX}} = \left\{ \overline{q_i^{TX}}[t] \right\}, i = 1,...,|\mathcal{U}_t| \tag{3.21}$$

9. **Average packet delay**: Based on Little's results [46] and those further extended in [47], [48], the average packet delay can be computed as follows:

$$\overline{d_i^{HoL}}[t] = \overline{q_i^{TX}}[t] / \overline{\lambda}_i[t] \tag{3.22}$$

The set of the average packet delay is described as:

$$\mathcal{S}_{9,t}^{S,C} = \mathcal{S}_t^{\overline{dHoL}} = \left\{ \overline{d_i^{HoL}}[t] \in \mathbb{R}_+ \right\}, i = 1,...,|\mathcal{U}_t| \tag{3.23}$$

The controllable subspace is computed based on the reunion of subsets from (1) to (9) as shown by Eq. 3.24:

$$\mathcal{S}_t^{S,C} = \bigcup_{p_j=1}^{9} \mathcal{S}_{p_j,t}^{S,C} \tag{3.24}$$

Based on the above parameters, the scheduler state space at TTI t can be represented as indicated by Eq. 3.25, where x_{i,p_i,p_j} are the controllable and uncontrollable parameters for each user $i \in \mathcal{U}_t$ and $\mathcal{S}_{i,t}^S$ is the user state.

$$\mathcal{S}_t^S = \bigcup_{p_i=1}^{6} \bigcup_{p_j=1}^{9} \bigcup_{i=1}^{|\mathcal{U}_t|} \mathsf{x}_{p_i,p_j,i} = \bigcup_{i=1}^{|\mathcal{U}_t|} \bigcup_{p_i=1}^{6} \bigcup_{p_j=1}^{9} \mathsf{x}_{i,p_i,p_j} = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{S}_{i,t}^S \tag{3.25}$$

The list of parameters being exposed in this sub-section represents part of the input parameters that different scheduling schemes consider in order to optimize the multi-objective problem. The rest of the parameters will be introduced in the following sections.

3.4 Radio Resource Allocation

The main focus of the LTE downlink scheduler is to assign the RBs with different instantaneous rates to different active users in order to satisfy given scheduling objectives. The idea is to quantify the benefit (utility) of allocating each RB $j \in \mathcal{B}$ to user $i \in \mathcal{U}_t$ at each TTI t. In this sense, the utility function has to be defined. In LTE scheduling, the utility functions cannot be measured directly. The solution is to perform the instantaneous rate allocation based on the utility representation at each TTI t and to measure or to evaluate the allocation performance at each TTI t+1 by using the objective functions. The instantaneous achievable rate matrix being obtained based on the CQI reports for each UE $i \in \mathcal{U}_t$ and for each RB $j \in \mathcal{B}$ is expressed in Eq. 3.26:

$$r = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,|\mathcal{B}|} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,|\mathcal{B}|} \\ \vdots & \vdots & & \vdots & & \vdots \\ r_{i,1} & r_{i,2} & \cdots & r_{i,j} & \cdots & r_{i,|\mathcal{B}|} \\ \vdots & \vdots & & \vdots & & \vdots \\ r_{|\mathcal{U}_t|-1,1} & r_{|\mathcal{U}_t|-1,2} & \cdots & r_{|\mathcal{U}_t|-1,j} & \cdots & r_{|\mathcal{U}_t|-1,|\mathcal{B}|} \\ r_{|\mathcal{U}_t|,1} & r_{|\mathcal{U}_t|,2} & \cdots & r_{|\mathcal{U}_t|,j} & \cdots & r_{|\mathcal{U}_t|,|\mathcal{B}|} \end{pmatrix} \in \mathbb{R}_{+}^{|\mathcal{U}_t| \times |\mathcal{B}|} \qquad (3.26)$$

where $r_{i,j}$ is the instantaneous achievable rate for user $i \in \mathcal{U}_t$ and RB $j \in \mathcal{B}$. The scheduling optimization problem refers to the rate allocation procedure from Eq. 3.26 in order to satisfy the scheduling objectives introduced in Chapter 1. The instantaneous achievable rate vector for a given LTE bandwidth is denoted by $r_i = \left[r_{i,1}\ r_{i,j} .... r_{i,|\mathcal{B}|} \right]$. Let us consider $U_i\left(r_i\right)$ the utility function which is a benefit representation of allocating the rates r_i to user $i \in \mathcal{U}_t$. The LTE scheduler aims to maximize the aggregate user utilities in the long term purpose such as:

$$\max_{t \to \infty} U\left(r\right) \qquad (3.27)$$

where $U\left(r\right) = 1 \Big/ \left(\left| \mathcal{U}_t \right| \cdot N_{TTI} \right) \cdot \sum_{t=1}^{N_{TTI}} \sum_{i=1}^{|\mathcal{U}_t|} U_i \left(r_i\left[t\right] \right)$ and $N_{TTI} \to \infty$ is the number of TTIs for a given downlink transmission.

If the RB allocation policy for the active users is $\pi_{RB} = \{b_{i,j}[t]\}, \forall i \in |\mathcal{U}_t|$, where $j = 1,...,|\mathcal{B}|$ and $t = 1,...,N_{TTI}$, then the decision matrix for the RB assignment at TTI t is $b[t] = \{b_{i,j}[t], i = 1,...,|\mathcal{U}_t|, j = 1,...,|\mathcal{B}|\}$ and takes the binary values as suggested by Eq. 3.28:

$$b_{i,j} = \begin{cases} 1, \ if \ \ RB \ \ j \in \mathcal{B} \ is \ allocated \ to \ UE \ i \in \mathcal{U}_t \\ 0, \ otherwise \end{cases} \tag{3.28}$$

The instantaneous data rates $R_i[t]$ for each user are obtained after performing the scheduling decision under the allocation policy π_{RB} at each TTI t. Let us define the instantaneous rate region constrained by policy π_{RB} such as $\mathcal{R}^R_{\pi_{RB}} \in \mathbb{R}_+^{|\mathcal{U}_t| \times |\mathcal{B}|}$. Therefore, the definition domain for the utility function is $U : \mathcal{R}^R_{\pi_{RB}} \to \mathbb{R}$ and the long-term optimization problem becomes [49], [50]:

$$\begin{aligned} & \max_{\pi_{RB}} \ \sum_{i=1}^{|\mathcal{U}|} U_i \left(\sum_{j=1}^{|\mathcal{B}|} b_{i,j} \cdot r_{i,j} \right) \\ & s.t. \ \ \sum_{i=1}^{|\mathcal{U}|} b_{i,j} = 1, \ \ j = 1,...,|\mathcal{B}| \\ & \qquad b_{i,j} \in \{0,1\}, \ \ \forall i \in \mathcal{U}, \forall j \in \mathcal{B} \end{aligned} \tag{3.29}$$

According to [51], [53], the local maximum is also a global maximum in Eq. 3.29 if and only if the region $\mathcal{R}^R_{\pi_{RB}}$ is a *convex set* and $U_i(R_i)$ is a *concave function*. However, the convexity problem of $\mathcal{R}^R_{\pi_{RB}}$ in OFDM systems has been discussed intensively in [50] and the authors came with the conclusion that the short-term optimization problem at each TTI t can be obtained by using the first order approximation of Taylor's expansion as expressed by Eq. 3.30 [50]:

$$\sum_{i=1}^{|\mathcal{U}|} U_i(R_i[t]) - \sum_{i=1}^{|\mathcal{U}|} U_i(R_i[t-1]) \approx \sum_{i=1}^{|\mathcal{U}|} U_i'(R_i[t-1]) \cdot (R_i[t] - R_i[t-1]) \tag{3.30}$$

where $U_i'(R_i[t-1]) = \partial U_i(R_i[t-1]) / \partial R_i[t-1]$ is the marginal utility for each user $i \in \mathcal{U}_t$. The instantaneous rate for each user $i \in \mathcal{U}_t$ $(R_i[t-1])$ at TTI t-1 is obtained after performing the scheduling procedure at TTI t-1 and this value is used in the

optimization problem at TTI *t*. Therefore, the short-term optimization problem can be written under the following form:

$$\begin{aligned} (P): & \max_{\pi_{RB}[t]} \sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|} b_{i,j}[t]\cdot U_i'\left(R_i[t]\right)\cdot r_{i,j}[t] \\ (C): & s.t. \quad \sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t]=1, \;\; j=1,...,|\mathcal{B}| \\ & b_{i,j}[t]\in\{0,1\}, \forall i\in\mathcal{U}_t, \forall j\in\mathcal{B} \end{aligned} \tag{3.31}$$

where (P) is the optimization problem, (C) represents the set of convex constraints of problem (P). The optimization model being exposed in Eq. 3.31 represents a *linear programming model* where the unknown variables are the resource assignment variables $b_{i,j}[t]\in\{0,1\}$ which have to be determined at each TTI *t* subject to set of constraints (C). Then, the idea is to find at each TTI *t* the optimal resource allocation policy $\pi_{RB}^{*}[t]$ which permits to select for the radio resource allocation, the set of users which is able to maximize the optimization problem (P). Due to the reduced number of RBs which has to be allocated at each TTI, the resource assignment is performed by using the following equation:

$$m_j[t] = \underset{i\in\mathcal{U}_t}{arg\,max}\left\{U_i'\left(R_i[t]\right)\cdot r_{i,j}[t]\right\} \tag{3.32}$$

where $m_j[t]$ indicates that RB $j\in\mathcal{B}$ is assigned or allocated to user $m\in\mathcal{U}_t$ $\forall m\neq i$ at TTI *t*. Consequently, $b_{m,j}[t]=1$ and $b_{i,j}[t]=0,\ \forall i\in\mathcal{U}_t$ and $\forall i\neq m$. This way, the user assignment is performed for each RB for a given LTE bandwidth. Once the resource allocation is finished, the TB size is determined for each selected user.

As mentioned earlier, the instantaneous rate for each user $i\in\mathcal{U}_t$ and for each RB $j\in\mathcal{B}$ is determined based on the CQI reports available at the CQI state space module. The marginal function is positive ($U_i':\mathcal{R}_{\pi_{RB}}^{R}\rightarrow\mathbb{R}^{+}$) because the utility function $U_i(R_i)$ must be concave (the second derivative is negative) in order to assure the linearity of the considered optimization problem. When the

utility function $U_i(R_i)$ takes the polynomial form, the role of its marginal utility is to schedule those users with the highest instantaneous rates by increasing at the same time the total system capacity if the radio channels are errorless. In the case of retransmissions, the MUF as a function of instantaneous user throughput $U_i'(T_i)$ should be used in order to provide more RBs to those users which require less retransmissions during the downlink scheduling session.

The linear programming model exposed in Eq. 3.31 is a typical SSR-SMOO problem being focused on the system throughput maximization without considering other objectives such as: user fairness, GBR, HoL delay, packet loss or stability requirements. The impact of the resource allocation problem in the scheduling objectives can be measured by using the objective functions. The objectives functions can be modeled by using the QoS constraints from Table 2.1. When different objective(s) is (are) analyzed, the performance of the optimization problem from Eq. 3.31 can be improved if the MU function considers the performance parameter(s) of the addressed objective(s). More details about this aspect are presented in the following section.

3.5 Utility and Objective Functions in LTE

Utility functions are designed to quantify the benefit of allocating a given and finite number of RBs to a number of active bearers. The type of utility function can influence the optimization problem in the direction of different scheduling objectives. The classification of utility functions can be achieved by considering three perspectives: the argument function, the utility weight and the manufacturing methodologies. Based on the manufacturing methods, there are two modes to obtaining the utility functions [49], [50] which are exposed bellow together with the proposed methodology:

- **Application based utility functions**: One way is to develop utility functions that characterize a specific type of application which can be obtained by using sophisticated subjective surveys. These utilities can suffer from the imperfection of the measurements, and different

parameters are fixed to some objective values denoting the inflexibility for those situations which are not covered by the considered surveys.

- **Traffic habits based utility functions**: statistics about the percentage of different traffic types which can exist at different moments of time in different urban scenarios. The utility functions are designed based on these statistics of heterogeneous traffic types.
- ***Scheduler state based aggregate utility function***: Based on a given scheduler state $\mathcal{S}_t^S$, different utility functions (which are already proposed in the specialty literature) are applied in order to maximize the long term aggregate utility function and to solve the DSR-SMOO/CMOO combinatorial problems. More precisely, it maximizes the sum of some existing utility functions subject to objective requirements. Details about this novel concept are highlighted in the following section.

The short-term optimization for the resource allocation is obtained when performing the first order of Taylor's expansion between two time consecutive utility functions. This way, the marginal utility function or the first derivative utility function is obtained. The term of marginal can refer also to a small change which can appear in the optimization problem between two consecutive scheduling procedures. In fact, the marginal utility indicates the obtained gain of scheduling objectives when performing the resource allocation procedure at each TTI. Therefore, more resources are to be allocated to those users with the highest gain in the marginal utility value. In the optimization problem being exposed in Eq. 3.31, when selecting any gain in the MU function leads to the system capacity maximization without any consideration about other objectives. By designing the marginal utility with proper weights, different objectives can be addressed. So, the role of the marginal utility in the optimization problem is to reduce the impact of the instantaneous achievable rate $r_{i,j}[t]$ (or to annihilate any variation of the radio channel) and to focus the entire optimization problem on scheduling different users which are unsatisfied from the viewpoint of the objective(s) which is (are) addressed by different MU weights. To conclude, the multi-objective performance depends on the type of marginal utility which is used in the optimization problem.

Let us define $\mathcal{X}_i$ the argument of the utility function for user $i \in \mathcal{U}_t$ and $\mathcal{Y}_i$ the argument for the utility weight, where $\mathcal{X}_i \cap \mathcal{Y}_i = \varnothing,\ \forall i \in \mathcal{U}_t$, $\mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{S}_t^S$, $\mathcal{X} = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{X}_i$ and $\mathcal{Y} = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{Y}_i$. Therefore the utility function for user $i \in \mathcal{U}_t$ can be decomposed as shown in Eq. 3.33:

$$U_i(\mathcal{X}_i) = W_i(\mathcal{Y}_i) \cdot F_i(\mathcal{X}_i) \tag{3.33}$$

where function F_i is concave and differentiable, $F_i : \mathcal{R}_{\pi_{RB}}^{\mathcal{X}} \to \mathbb{R}$ and $W_i : \mathcal{R}_{\pi_{RB}}^{\mathcal{Y}} \to \mathbb{R}$, where $\mathcal{R}_{\pi_{RB}}^{\mathcal{X}}$ and $\mathcal{R}_{\pi_{RB}}^{\mathcal{Y}}$ are the regions of performance parameters $\mathcal{X} \subseteq \mathcal{S}_t^S$ and $\mathcal{Y} \subseteq \mathcal{S}_t^S$, respectively, constrained by the allocation policy π_{RB}. The utility weight $W_i(\mathcal{Y}_i)$ of user $i \in \mathcal{U}_t$ is a constant, but it is represented as a function in order to highlight the objective parameter $\mathcal{Y}_i$, where $\mathcal{Y}_i$ can be the HoL packet delay, the user throughput, the packet loss rate, the average queue size, etc. When the weight argument $\mathcal{Y}_i$ respects different QoS or objective constraints, then user $i \in \mathcal{U}_t$ is satisfied from the viewpoint of the objective addressed by $\mathcal{Y}_i$. The first derivative for the utility function is determined by using the following relation: $U_i'(\mathcal{X}_i) = W_i(\mathcal{Y}_i) \cdot F_i'(\mathcal{X}_i)$, where $F_i'(\mathcal{X}_i) = \partial F_i(\mathcal{X}_i) / \partial \mathcal{X}_i$. If the MU function is developed in such a way that the radio channel variations are compensated at each TTI t for each user $i \in \mathcal{U}_t$ $\left(r_{i,j}[t] \cdot F_i'(\mathcal{X}_i[t]) \approx 1\right)$, then the optimization problem is focused more on the scheduling objective evaluated by the weight argument $\mathcal{Y}_i$.

3.5.1 SSR Based SMOO/CMOO Problems

The short-term optimization can be obtained for the general form of scheduling utilities by using the first order of Taylor's expansion and being similar to Eq. 3.30 such as [50]:

$$\sum_{i=1}^{|\mathcal{U}_t|} U_i(\mathcal{X}_i[t]) - \sum_{i=1}^{|\mathcal{U}_t|} U_i(\mathcal{X}_i[t-1]) \approx \sum_{i=1}^{|\mathcal{U}_t|} W_i(\mathcal{Y}_i[t]) \cdot F_i'(\mathcal{X}_i[t]) \cdot (R_i[t] - R_i[t-1]) \tag{3.34}$$

where the utility argument can be $\mathcal{X}_i \in \left\{R_i, \overline{T}_i, \overline{d_i^{HoL}}\right\}$. More details about the utility

functions for different objectives are presented in the upcoming sub-sections and in Section 3.8 where the relevant related work in LTE scheduling is analyzed.

Let us consider the objective index $o=1,...,|\mathcal{O}|$, where $\mathcal{O}$ represents the set of scheduling objectives in LTE. For each objective $o \in \mathcal{O}$, let us define the pool of utilities $\mathcal{PU}_o$, and then, the entire set of utilities for all objectives is defined as $\mathcal{PU} = \bigcup_{o=1}^{|\mathcal{O}|} \mathcal{PU}_o$. As mentioned earlier, the type of marginal utility for objective $o \in \mathcal{O}$ is correlated with the utility weight. If $w_o = 1,...,|\mathcal{PU}_o|$ is the weight index targeting the objective $o \in \mathcal{O}$ and if $|\mathcal{PU}_o|$ is the number of utility weights for the objective $o \in \mathcal{O}$, then the generalized optimization problem for the same objective $o \in \mathcal{O}$ with the marginal utility weight $w_0 \in \mathcal{PU}_o$ becomes:

$$\begin{aligned} &(P_o): \ \max_{\pi_{RB}[t]} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} b_{i,j}[t] \cdot W^o_{w_o,i}\left(\mathcal{Y}^o_{w_o,i}[t]\right) \cdot F'^o_{w_o,i}\left(\mathcal{X}^o_{w_o,i}[t]\right) \cdot r_{i,j}[t] \\ &(C_o): \ s.t. \ \sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t] = 1, \ j = 1,...,|\mathcal{B}| \\ &\qquad b_{i,j}[t] = \{0,1\}, \ \forall i \in \mathcal{U}_t, \forall j \in \mathcal{B} \end{aligned} \tag{3.35}$$

where (P_o) is the optimization problem focusing on the objective $o \in \mathcal{O}$, (C_o) is the convex set of constraints for the objective $o \in \mathcal{O}$ and $W^o_{w_o,i}$ is the marginal utility weight $w_0 \in \mathcal{PU}_o$ of user $i \in \mathcal{U}_t$ for objective $o \in \mathcal{O}$. The optimization problem of Eq. 3.35 is a linear programming model. If $W^o_{w_o,i}[t] \gg F'^o_{w_o,i}[t] \cdot r_{i,j}[t]$, then the addressed objective of problem (P_o) is evaluated based on the weight argument $\mathcal{Y}_i$ for each user $i \in \mathcal{U}_t$. Let us define the weight matrix for objective $o \in \mathcal{O}$ such as: $\nabla U^o = \left\{U'^o_{w_o,i} = W^o_{w_o,i} \cdot F'^o_{w_o,i}, w_o = 1,...,|\mathcal{PU}_o|, i = 1,...,|\mathcal{U}_t|\right\}$. In the case of SSR-SMOO problems, the weight matrix ∇U^o assigns the same type of marginal utility functions to each user $i \in \mathcal{U}_t$ for the entire scheduling session.

The optimal resource allocation when following the linear optimization problem (P_o) and the set of constraints (C_o) for each RB $j \in \mathcal{B}$ and for a group of selected users $\forall i \in \mathcal{U}_t$ is given by the assignment being illustrated in Eq. 3.36:

$$m_j[t] = \underset{i \in \mathcal{U}_t}{arg\,max} \left\{ W^o_{w_o,i}\left(\mathcal{Y}^o_{w_o,i}[t]\right) \cdot F'^o_{w_o,i}\left(\mathcal{X}^o_{w_o,i}[t]\right) \cdot r_{i,j}[t] \right\} \tag{3.36}$$

where $m_j[t]$ indicates that RB $j \in \mathcal{B}$ is allocated to user $m \in \mathcal{U}_t$, $\forall m \neq i$ at TTI *t*. Then, $b_{m,j}[t] = 1$ and $b_{i,j}[t] = 0, \forall m \in \mathcal{U}_t, \forall i \in \mathcal{U}_t$ and $\forall i \neq m$. Therefore, the scheduling rule function can be defined in the following manner:

$$D^o_{w_o,i} : \mathbb{R} \to \mathbb{R} \quad D^o_{w_o,i}\left(\mathcal{J}^o_{w_o,i}\right) = W^o_{w_o,i}\left(\mathcal{Y}^o_{w_o,i}\right) \cdot F'^o_{w_o,i}\left(\mathcal{X}^o_{w_o,i}\right) \cdot r_{i,j} \tag{3.37}$$

where $\mathcal{J}^o_{w_o,i} = \mathcal{X}^o_{w_o,i}$ if $W^o_{w_o,i}\left(\mathcal{Y}^o_{w_o,i}\right) = 1$ and $\mathcal{J}^o_{w_o,i} = \mathcal{Y}^o_{w_o,i}$, otherwise. If the objective $o \in \mathcal{O}$ and the utility weight $w_0 \in \mathcal{PU}_o$ remain fixed during the entire scheduling session, then the linear optimization problem (P_o) is entitled SSR-SMOO/CMOO problem. The CMOO refers to the fact that $\mathcal{Y}^o_{w_o,i}$ can be multi-dimensional and the optimization problem is focusing on the multi-objective criteria.

The performance of the scheduling discipline $D^o_{w_o,i}$ after the resource allocation procedure can be evaluated by using the objective functions. Let us define the objective function $\phi_{o,i}\left(\mathcal{J}^o_{w_o,i}\right)$ for objective $o \in \mathcal{O}$ and user $i \in \mathcal{U}_t$, where the definition domain is $\phi_{o,i} : \mathbb{R} \to \mathbb{R}$. The objective condition for a given SSR-SMOO problem for each user $i \in \mathcal{U}_t$ at each TTI is given by Eq. 3.38:

$$(O_o): \ \phi_{o,i}\left(\mathcal{J}^o_{w_o,i}[t]\right) \geq 0 \ , \forall o \in \mathcal{O}, \forall i \in \mathcal{U}_t \tag{3.38}$$

When the condition is satisfied for each user $i \in \mathcal{U}_t$, then the scheduler becomes optimal at TTI *t* from the viewpoint of objective $o \in \mathcal{O}$. Therefore, the aggregated function for the entire set of active users and objective $\forall o \in \mathcal{O}$ becomes: $\phi_o\left(\mathcal{J}^o_{w_o}[t]\right) = \left(1/|\mathcal{U}_t|\right) \cdot \sum_{i=1}^{|\mathcal{U}_t|} \phi_{o,i}\left(\mathcal{J}^o_{w_o,i}[t]\right)$, where $\phi_o : \mathbb{R} \to \mathbb{R}$ and $\mathcal{J}^o_{w_o} = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{J}^o_{w_o,i}$.

In the case of DSR-CMOO problems, the impact of each scheduling rule in the aggregate functions for each objective is strongly required. In this sense, the aggregate multi-objective function $\Phi_{o,w_o} : \mathbb{R} \to \mathbb{R}$ when applying the scheduling discipline $D^o_{w_o,i}$, can be represented as indicated in Equation 3.39:

$$\Phi_{o,w_o}\left(\mathcal{J}^{o}_{w_o}[t]\right)=\sum_{o^*=1}^{|\mathcal{O}|}\delta_{o^*}\cdot\phi^{o,w_o}_{o^*}\left(\mathcal{J}^{o^*}_{w_{o^*}}[t]\right) \tag{3.39}$$

where $\phi^{o,w_o}_{o^*}:\mathbb{R}\rightarrow\mathbb{R}$ is the aggregate function of objective $o^*\in\mathcal{O}$ when the scheduling rule $D^{o}_{w_o,i}$ is applied $\forall o\in\mathcal{O}$, $\forall w_o\in\mathcal{PU}_o$ for each user $i\in\mathcal{U}_t$. The parameter δ_{o^*} is the weight for the particular objective function $\phi^{o,w_o}_{o^*}$, $\forall o^*\in\mathcal{O}$. The necessary condition to be satisfied by the aggregate multi-objective function at each TTI *t* for the entire set of active users is highlighted in Eq. 3.40:

$$\Phi_{o,w_o}\left(\mathcal{J}^{o}_{w_o}[t]\right)\geq 0 \tag{3.40}$$

Equation 3.40 should be satisfied only and only if conditions from Eq. 3.38 (O_o) are met for each user $i\in\mathcal{U}_t$ and for each objective $o\in\mathcal{O}$. Precise details about the particular objective functions are provided in the following sub-sections.

3.5.2 Utility and Objective Functions for Throughput Maximization

When the throughput maximization objective ($o=1$) is considered, the utility argument is $\mathcal{X}^{1}_{w_1,i}=R_i$ and the weight function is $W^{1}_{w_1,i}\left(\mathcal{Y}^{1}_{w_1,i}\right)=1$. The MOO evaluator grants the scheduling performance based on the user objective function $\phi_{1,i}:\mathbb{R}\rightarrow\mathbb{R}$, $\forall i\in\mathcal{U}_t$ where $\phi_{1,i}[t]=\sum_{i=1}^{|\mathcal{U}_t|}T_i[t]-\sum_{i=1}^{|\mathcal{U}_t|}T_i[t-1]$. The role of the optimization problem is to increase the total cell throughput TTI-by-TTI such that $\phi_{1,i}[t]\geq 0$. If the MUF from Eq. 3.31 is $U'^{1}_{1,i}\left(R_i[t]\right)=1$, then the obtained scheduling rule is entitled ***Maximum Throughput*** (MT), aiming to maximize for each TTI the total cell spectral efficiency.

3.5.3 Utility and Objective Functions for User Fairness

By adopting the optimization problem (P_1) focusing on the capacity maximization under the fairness requirement, the fairness should be guaranteed at each TTI *t*, degrading at the same time the spectral efficiency performance.

Therefore, a new scheme is required in order to give more flexibility to the system throughput improvement.

The time window (a given number of TTIs) constrains or relaxes the fairness performance depending on its length. By adopting the average user throughput from Eq. 3.18 as an argument for the MUF, the resource allocation at TTI *t* depends on the allocation history in the previous TTIs. The averaging procedure can be achieved in two ways:

1. By using the exponential moving filter, the obtained throughput is entitled Average User Throughput with Exponential Moving Filter (AUT-EMF): The forgetting factor $\beta_{\bar{T}}$ from Eq. 3.18 is used to control the system throughput and user fairness tradeoff, where $\beta_{\bar{T}} = 1_{TTI} / T_w^E$, and T_w^E is the time window length. This means that a higher average throughput implies a lower priority for that UE to be selected on the considered RB. The only condition is to set $\beta_{\bar{T}}$ larger than the channel correlation time in order to exploit the time diversity principle [50]. It is important to note that if the time window T_w^E is too large, the cell spectral efficiency is affected and if the time window is too small, then the user fairness is not sensed anymore.
2. By using the median moving filter, the obtained throughput is entitled Average User Throughput with Median Moving Filter (AUT-MMF). The idea is to store the instantaneous user throughputs for a given time window T_w^M and to use the mean value of these observations at each TTI in order to balance the system throughput and user fairness tradeoff.

For different reasons which are explained in Chapter 6, both types of observations are used in the scheduling procedure: AUT-EMF computes the scheduling rule and AUT-MMF determines the objective function. Without going through precise details at this stage, it can be specified that the NGMN fairness criterion is used in order to compute the objective function [52]. Based on this principle, the Cumulative Distribution Function (CDF) calculated for a given set of average/instantaneous user throughputs should not exceed a given NGMN threshold. More details about this concept are provided in Chapter 6.

Based on some convergence studies presented in [50], the local optimization considers $\mathcal{X}^2_{w_2,i} = \overline{T}_i[t]$ as an argument for the utility function and the weight function is $W^2_{w_2,i}\left(\mathcal{Y}^2_{w_2,i}\right) = 1$ since the QoS requirements are not included in the utility function.

A particular type of marginal utility function is $U'^2_{1,i}\left(\overline{T}_i[t]\right) = 1/\overline{T}_i[t]$ which implies the metric of $D^2_{1,i}\left(\overline{T}_i[t]\right) = r_{i,j}[t]/\overline{T}_i[t]$, known in the literature as the ***Proportional Fair*** (PF) scheduling rule [53]. By adopting PF as a static scheduling rule, a certain degree of tradeoff between user fairness and system throughput can be achieved. When $U^2_{w_2,i}\left(\overline{T}_i[t]\right) = \overline{T}_i[t]$, the obtained scheduling rule becomes MT. Therefore, by setting various forms of the PF utility functions, different levels of fairness are obtained.

The NGMN objective function which is considered in this particular case is $\phi_{2,i}[t] = \psi_i^{Req}\left[\widehat{T}_i\right][t] - \psi_i\left[\widehat{T}_i\right][t]$, where $\widehat{T}_i$ is the Normalized User Throughput (NUT), ψ_i is the CDF function for a given distribution of NUT observations and ψ_i^{Req} represents the NGMN fairness requirement [52]. More details about the system model under NGMN fairness constraints are provided in Chapter 6 where $\widehat{T}_i$ is modeled based on both AUT-MMF and AUT-EMF observations.

3.5.4 Utility and Objective Functions for Guaranteed User Throughput

When the rate constraint satisfaction ($o = 3$) is considered in the optimization problem, the utility function weight should be aware of the newest parameters such as $\mathcal{X}^3_{w_3,i} = \overline{T}_i[t]$ and $\mathcal{Y}^3_{w_3,i} = \left\{\overline{\overline{T}}_i[t]\right\}$, where $\overline{\overline{T}}_i[t]$ represents the average user throughput calculated by using the median filter. The objective function for the GBR satisfaction can be formulated as $\phi_{3,i}[t] = \overline{\overline{T}}_i[t] - \underline{T}_i[t]$ and the objective condition imposes that the mean user throughput should be greater

than the GBR requirement at each TTI *t*. The optimization function focusing on the MBR requirement is not covered in this study.

In particular, Equation 3.41 represents a typical example which belongs to the class of GBR utilities where $D_{1,i}^{3}\left(\overline{\overline{T}}_i[t]\right)$ is the scheduling rule known as Barrier Function based PF (BF-PF) [54].

$$\begin{cases} W_{1,i}^{3}\left(\overline{\overline{T}}_i[t]\right) = 1 + \omega_{1,1}^{3} \cdot exp\left[-\omega_{2,1}^{3} \cdot \left(\overline{\overline{T}}_i[t] - \underline{\overline{T}}_i[t]\right)\right] \\ F_{1,i}^{'3}\left(\overline{T}_i[t]\right) = 1/\overline{T}_i[t] \\ D_{1,i}^{3}\left(\overline{\overline{T}}_i[t]\right) = \left\{1 + \omega_{1,1}^{3} \cdot exp\left[-\omega_{2,1}^{3} \cdot \left(\overline{\overline{T}}_i[t] - \underline{\overline{T}}_i[t]\right)\right]\right\} \cdot \dfrac{r_{i,j}[t]}{\overline{T}_i} \end{cases} \tag{3.41}$$

By applying the BF-PF rule at each TTI, data flows which are not able to guarantee the minimum bit rate from the perspective of AUT-MMF observations are preferred to be scheduled in the current time instant. When all bearers are satisfied from the GBR constraint point of view, the fairness maintenance becomes the main objective.

3.5.5 Utility and Objective Functions for HoL Packet Delay

If the utility weight depends on the instantaneous HoL delay ($o = 4$) such that $\mathcal{Y}_{w_4,i}^{4} = \left\{d_i^{HoL}[t]\right\}$ and $\mathcal{X}_{w_4,i}^{4} = \overline{T}_i[t]$, then the optimization problem considers the HoL packet delay as the first priority in the satisfaction of the performance criterion. Then, the delay based objective function which has to be maximized for each active data queues at each TTI *t* becomes $\phi_{4,i}[t] = \overline{d_i^{HoL}}[t] - d_i^{HoL}[t]$.

For the particular case of Equation 3.42, the resulted scheduling discipline is the well-known Modified Largest Weighted Delay First (M-LWDF) [55]. The parameter $\omega_{1,i}^{4}$ differentiates the real-time traffic based on the delay and PLR constraints. Obviously, M-LWDF scheme prefers the flows with larger HoL delays to be scheduled at each TTI.

$$\begin{cases} W_{1,i}^{4}\left(d_i^{HoL}\left[t\right]\right)=\omega_{1,i}^{4}\cdot d_i^{HoL}\left[t\right] \\ F_{1,i}^{'4}\left(\overline{T}_i\left[t\right]\right)=1/\overline{T}_i\left[t\right] \\ D_{1,i}^{4}\left(d_i^{HoL}\left[t\right]\right)=\omega_{1,i}^{4}\cdot d_i^{HoL}\left[t\right]\cdot r_{i,j}\left[t\right]/\overline{T}_i\left[t\right] \\ \omega_{1,i}^{4}=-log\left(\overline{R}_i^{PL}\left[t\right]\right)/\overline{d}_i^{HoL}\left[t\right] \end{cases} \tag{3.42}$$

For the particular case when all the active flows experiencing the same delays and packet loss rates, the M-LWDF rule acts as a pure PF scheduling rule, assuring fairness-throughput tradeoff levels based on the $\mathcal{S}_t^{CQI}$ conditions. More details about the M-LWDF scheduling rule and its integration into DSR-CMOO problems will be given in Chapter 7.

In general, the packet dropping module and scheduling procedures focusing on the HoL delay work in close collaboration. When the packet exceeds a given HoL constraint, the packet is automatically dropped. It is very interesting to study how these two objectives work under the DSR-CMOO optimization problem. Details about this concept are presented in Chapter 7.

3.5.6 Utility and Objective Functions for Packet Loss

The PLR objective ($o=5$) represents an important performance target in LTE scheduling. The utility weight depends on $\mathcal{Y}_{w_5,i}^{5}\left[t\right]=\left\{R_i^{PL}\left[t\right]\right\}$ and the utility argument keeps a similar form of $\mathcal{X}_{w_5,i}^{5}=\overline{T}_i\left[t\right]$. The objective function used to measure the performance of radio resource allocation from the viewpoint of PLR becomes: $\phi_{5,i}\left[t\right]=\overline{R}_i^{PL}\left[t\right]-R_i^{PL}\left[t\right]$. The way the PLR rate $R_i^{PL}\left[t\right]$ is computed is very important in the objective satisfaction. For instance, in Chapter 7 the same median filter time window is used for the PDR rate computation.

Let us consider a specific case of utility functions being composed of $F_{1,i}^{5}\left(\overline{T}_i\left[t\right]\right)$ and $W_{1,i}^{5}\left(R_i^{PL}\left[t\right]\right)$, as shown in Eq. 3.43. The obtained scheduling rule $D_{1,i}^{5}\left(R_i^{PL}\left[t\right]\right)$ is entitled the Packet Loss Fair based PF (PLF-PF) [56]:

$$\begin{cases} W_{1,i}^{5}\left(R_i^{PL}[t]\right) = R_i^{PL}[t] \\ F_{1,i}^{5}\left(\overline{T}_i[t]\right) = log\left(\overline{T}_i[t]\right) \\ D_{1,i}^{5}\left(R_i^{PL}[t]\right) = R_i^{PL}[t]\cdot r_{i,j}[t] / \overline{T}_i[t] \end{cases} \quad (3.43)$$

As expected, PLF-PF allocates more resources to such users with higher PLRs. When users experience appropriate PLRs performance, PLF-PF performs as a pure PF rule improving at the same time the fairness performance.

Other interesting scheduling rule is the Opportunistic Packet Loss Fair based PF (OPLF-PF) [56] which represents in fact the SSR-CMOO problem being focused on HoL packet delay and PLR objectives as indicated in Eq. 3.44:

$$\begin{cases} W_{2,i}^{5}\left(R_i^{PL}[t]\right) = R_i^{PL}[t]\cdot d_i^{HoL}[t] \Big/ \overline{R}_i^{PL}[t]\cdot \overline{d}_i^{HoL}[t] \\ F_{2,i}^{5}\left(\overline{T}_i[t]\right) = log\left(\overline{T}_i[t]\right) \\ D_{2,i}^{5}\left(R_i^{PL}[t]\right) = R_i^{PL}[t]\cdot d_i^{HoL}[t] \Big/ \overline{R}_i^{PL}[t]\cdot \overline{d}_i^{HoL}[t]\cdot r_{i,j}[t] / \overline{T}_i[t] \end{cases} \quad (3.44)$$

When the PLR and HoL delay objectives are satisfied for each user $i \in \mathcal{U}_t$, the scheduler is considered feasible from the viewpoint of the aforementioned objectives. Details about DSR-CMOO problems focusing on PDR and HoL delay multi-objective criterion are provided in Chapter 7.

3.5.7 Utility and Objective Functions for Queue Stability

The optimization problems considered so far aim to maximize the long term aggregate utility function in terms of $\overline{T}_i[t]$. When the global optimization considers the average HoL delay $\mathcal{X}_{w_6,i}^{6} = \left\{\overline{d_i^{HoL}}\right\}$ as the utility argument, the long-term optimization is $\underset{i\in\mathcal{U}_t}{max}\left\{1/\left(|\mathcal{U}_t|\cdot N_{TTI}\right)\cdot\sum_{t=1}^{N_{TTI}}\sum_{i=1}^{|\mathcal{U}_t|}U_i\left(\overline{d_i^{HoL}}[t]\right)\right\}$. The short term optimization problem becomes tractable by using the first order approximation of Taylor's expansion [50]. In order to address the queue stability objective ($o = 6$), the utility weight has to compensate the rate $r_{i,j}[t]$ variations by simply setting the weight argument to $\mathcal{Y}_{w_6,i}^{6}[t] = \left\{\overline{T}_i[t]\right\}$.

The objective function which has to be considered by each active data queue is defined as $\phi_{6,i}[t] = \overline{q_i^{TX}}[t] / 1_{TTI} - \overline{T_i}[t]$ where 1_{TTI} is the duration of one TTI. Based on the objective condition, the system should be aware of whether in the next TTI the supply of each queue is sufficient to support a new transmission without any waste of the radio resources.

For the particular case exposed in Eq. 3.45, the corresponding scheduling discipline $D_{1,i}^{6}\left(\overline{q_i^{TX}}[t]\right)$ is entitled Max-Delay Utility based PF (MDU-PF) for the BE traffic type, initially proposed in [49] and [50] . This scheduling rule is used in the optimization problem in Chapter 7 which considers the DSR-CMOO problem focusing on NGMN fairness, GBR, HoL and PDR objectives.

$$\begin{cases} W_{1,i}^{6}\left(\overline{T}_i[t]\right) = \overline{\lambda_i}[t] / \overline{T}_i[t] \\ F_{1,i}^{6}\left(\overline{d_i^{HoL}}[t]\right) = \left(\overline{d_i^{HoL}}[t]\right)^2 / 2 \\ D_{1,i}^{6}\left(\overline{q_i^{TX}}[t]\right) = \overline{q_i^{TX}}[t] \cdot r_{i,j}[t] / \overline{T}_i[t] \end{cases} \tag{3.45}$$

The MDU-PF serves each queue with a larger length, assuring at the same time the system stability. When all active flows experience relatively equal queue lengths, the proportional fairness assurance becomes the first objective.

3.6 Aggregate Utility Based MOO Problem

The optimization problems $(P_o), \forall o \in \mathcal{O}$ address the SMOO scheduling technique depending on the considered criterion. Each SMOO problem is considered to be linear guaranteeing at the same time the global optimal solution. By adopting different utility functions, the scheduling procedure impacts differently in the MOO problem. Each objective from MOO defines its own pool of utilities. Each pool contains different utilities which target the same objective. The aggregate utility function is obtained as follows:

$$U^{Agg}(\mathcal{X}) = \frac{1}{|\mathcal{PU}| \cdot |\mathcal{U}_t|} \cdot \sum_{o=1}^{|\mathcal{O}|} \sum_{w_o=1}^{|\mathcal{PU}_o|} \sum_{i=1}^{|\mathcal{U}_t|} U_{w_o,i}^{o}\left(\mathcal{X}_{w_o,i}^{o}\right) \tag{3.46}$$

where $\mathcal{X} = \bigcup_{o,w_o,i} \mathcal{X}^{o}_{w_o,i}$, $o = 1,..,|\mathcal{O}|$, $w_o = 1,..,|\mathcal{PU}_o|$, $i = 1,..,|\mathcal{U}_t|$ represents the set of utility arguments for different objectives and for different active users. The proposed MOO problem in the long term purpose aims to maximize the sum of utilities for each scheduling objective as indicated by Eq. 3.47:

$$\max_{\mathcal{X} \subset S^S} \left[\frac{1}{N_{TTI}} \cdot \sum_{t=1}^{N_{TTI}} U^{Agg}\left(\mathcal{X}[t]\right) \right] \tag{3.47}$$

By decomposing $U^{Agg}(\mathcal{X})$ in sums of utility functions for each objective and by considering the arguments for the utility and MU functions, then Equation 3.47 can be decomposed as follows:

$$\begin{aligned} \max_{r[t] \in \mathcal{R}^R_{\pi_{RB}}} \Bigg[& \sum_{w_1=1}^{|\mathcal{PU}_1|} \sum_{i=1}^{|\mathcal{U}_t|} U^1_{w_1,i}(R_i) + \sum_{w_2=1}^{|\mathcal{PU}_2|} \sum_{i=1}^{|\mathcal{U}_t|} U^2_{w_2,i}\left(\overline{T}_i[t]\right) + \sum_{w_3=1}^{|\mathcal{PU}_3|} \sum_{i=1}^{|\mathcal{U}_t|} U^3_{w_3,i}\left(\overline{T}_i[t]\right) + \\ & \sum_{w_4=1}^{|\mathcal{PU}_4|} \sum_{i=1}^{|\mathcal{U}_t|} U^4_{w_4,i}\left(\overline{T}_i[t]\right) + \sum_{w_5=1}^{|\mathcal{PU}_5|} \sum_{i=1}^{|\mathcal{U}_t|} U^5_{w_5,i}\left(\overline{T}_i[t]\right) + \sum_{w_6=1}^{|\mathcal{PU}_6|} \sum_{i=1}^{|\mathcal{U}_t|} U^6_{w_6,i}\left(\overline{d_i^{HoL}}[t]\right) \Bigg] \end{aligned} \tag{3.48}$$

Each sum of user utilities from Eq. 3.48 represents a concave aggregate function which implicitly involves the concavity of the entire optimization problem. The set $\mathcal{R}^R_{\pi_{RB}}$ of the instantaneous user rates under a given policy π_{RB} of RB allocation is considered to be *convex*. Therefore, the global maximum is also a local maximum. By adopting the first order approximation of Taylor's expansion for each of the utility functions, the instantaneous optimization problem becomes:

$$\begin{aligned} \max_{\substack{r[t] \in \mathcal{R}^R_{\pi_{RB}} \\ w_o \in \mathcal{PU}_o}} \Bigg[& \sum_{w_1=1}^{|\mathcal{PU}_1|} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} F'^1_{w_1,i}\left(R_i[t]\right) \cdot r_{i,j}[t] + \sum_{w_2=1}^{|\mathcal{PU}_2|} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} F'^{2,}_{w_2,i}\left(\overline{T}_i[t]\right) \cdot r_{i,j}[t] + \\ & \sum_{w_3=1}^{|\mathcal{PU}_3|} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} W^3_{w_3,i}\left(\overline{\overline{T}_i}[t]\right) \cdot F'^3_{w_3,i}\left(\overline{T}_i[t]\right) \cdot r_{i,j}[t] + \\ & \sum_{w_4=1}^{|\mathcal{PU}_4|} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} W^4_{w_4,i}\left(d_i^{HoL}[t]\right) \cdot F'^4_{w_4,i}\left(\overline{T}_i[t]\right) \cdot r_{i,j}[t] + \\ & \sum_{w_5=1}^{|\mathcal{PU}_5|} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} W^5_{w_5,i}\left(R_i^{PL}[t]\right) \cdot F'^5_{w_5,i}\left(\overline{T}_i[t]\right) \cdot r_{i,j}[t] + \\ & \sum_{w_6=1}^{|\mathcal{PU}_6|} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} W^6_{w_6,i}\left(\overline{T}_i[t]\right) \cdot F'^6_{w_6,i}\left(\overline{d_i^{HoL}}[t]\right) \cdot r_{i,j}[t] \Bigg] \end{aligned} \tag{3.49}$$

By compressing the whole MOO problem, Eq. 3.49 is equivalent with Eq. 3.50:

$$\max_{\substack{r[t]\in\mathcal{R}^{R}_{\pi_{RB}} \\ o,w_o\in\mathcal{PU}_o}} \left[\sum_{o=1}^{|\mathcal{O}|} \sum_{w_o=1}^{|\mathcal{PU}_o|} \sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} W^{o}_{w_o,i}\left(\mathcal{Y}^{o}_{w_o,i}[t]\right)\cdot F'^{o}_{w_o,i}\left(\mathcal{X}^{o}_{w_o,i}[t]\right)\cdot r_{i,j}[t] \right] \tag{3.50}$$

Let us define the policy $\pi_{\nabla U}$ of selecting different scheduling rules from the pool of utilities (or rules) $\mathcal{PU}$. The Scheduling Rule Selection (SRS) policy acts similarly to the selection policy of RBs with the amendment that instead of users, $\pi_{\nabla U}$ considers the number of objectives and instead of RBs, the SRS policy takes into account the existing scheduling rules for each objective class. Another difference is the fact that at each TTI only one objective is required while multiple users can be scheduled within one TTI when π_{RB} is used. Therefore, the policy of selecting different objectives and marginal utilities at each TTI *t* is defined such as $\pi_{\nabla U}[t]=\left\{c_{o,w_o}[t]\right\}_{\forall o,w_o}$, where $t=1,...,N_{TTI}$ and $c_{o,w_o}[t]$ is the decision variable $\forall o\in\mathcal{O}$, $\forall w_o\in\mathcal{PU}_o$ and at TTI t $c[t]=\left\{c_{o,w_o}[t], o=1,...,|\mathcal{O}|, w_o=1,...,|\mathcal{PU}_o|\right\}$ is the decision matrix of scheduling rule. Based on Eq. 3.50, for each active user $i\in\mathcal{U}_t$, the same marginal utility $w_o\in\mathcal{PU}_o$ must be assigned at each TTI *t*. Then, the decision variable $u^{o}_{w_o,i}[t]$ who assigns MU functions $w_o\in\mathcal{PU}_o$ for objective $o\in\mathcal{O}$ to each user $i\in\mathcal{U}_t$ becomes mandatory in the short-term optimization problem. In this sense, the matrix $u^{o}[t]=\left\{u^{o}_{w_o,i}[t], w_o=1,...,|\mathcal{PU}_o|, i=1,...,|\mathcal{U}_t|\right\}$ assigns the same scheduling rule for objective $\forall o\in\mathcal{O}$ to each user $i\in\mathcal{U}_t$ at TTI *t* and this matrix differs from one TTI to another in the DSR-CMOO problems.

The aggregate MOO optimization problem based on $\pi_{\nabla U}$ and π_{RB} can be formulated as indicated by Eq. 3.51, where $\left(P^{Agg}_{Primal}\right)$ indicates the aggregate optimization problem and $\left(C^{Agg}_{Primal}\right)$ is the set of constraints. The first constraint denotes the necessary condition of selecting at each TTI *t* only one scheduling rule and $c_{o,w_o}[t]\in\{0,1\}$. The set of constraints (b) indicates that only one MU function is selected for the entire set of active users at each TTI *t*. Constraints (c)

$$\left(P_{Primal}^{Agg}\right):\ \max_{\pi_{RB},\pi_{\nabla U}} \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} c_{o,w_o}[t]\cdot\sum_{i=1}^{|\mathcal{U}_t|} u^{o}_{w_o,i}[t]\cdot\sum_{j=1}^{|\mathcal{B}|} b_{i,j}[t]\cdot U'^{o}_{w_o,i}\left(\mathcal{X}^{o}_{w_o,i}[t]\right)\cdot r_{i,j}[t]$$

$$\left(C_{Primal}^{Agg}\right):\ s.t.\quad \begin{aligned}
&\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} c_{o,w_o}[t]=1 && (a)\\
&\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} u^{o}_{w_o,i}[t]=1,\ i=1,...,|\mathcal{U}_t| && (b)\\
&\sum_{i=1}^{|\mathcal{U}_t|} u^{o}_{w_o^*,i}[t]=|\mathcal{U}_t|,\ w_o^*\in\mathcal{PU}_o, o\in\mathcal{O} && (c)\\
&\sum_{i=1}^{|\mathcal{U}_t|} u^{o}_{w_o^{\otimes},i}[t]=0,\ w_o^{\otimes}=1,...,|\mathcal{PU}_o|, o=1,...,|\mathcal{O}|, \forall w_o^{\otimes}\neq w_o^* && (d)\\
&\sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t]=1,\ j=1,...,|\mathcal{B}| && (e)\\
&c_{o,w_o}[t]\cdot\Phi_{o,w_o}[t+1]\geq 0,\ o=1,...,|\mathcal{O}|, w_o=1,...,|\mathcal{PU}_o| && (f)\\
&c_{o,w_o}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o &&\\
&u^{o}_{w_o,i}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t && (3.51)\\
&b_{i,j}[t]\in\{0,1\},\ \forall i\in\mathcal{U}_t,\forall j\in\mathcal{B} &&
\end{aligned}$$

and (d) assure that the same marginal utility function is assigned to all users at each TTI. The set of constraints (e) is the well-known set of conditions of assigning the RBs to different users. Finally, the set of constraints (f) considers the aggregate multi-objective condition from Eq. 3.40. This implies that for the selected rule $\left(c_{o,w_o}[t]\right)$, the sum of aggregate functions for each objective at TTI *t*+1 should be greater than zero since the evaluation of the scheduling procedure is performed in the next TTI. If the objective conditions from Eq. 3.38 are satisfied, then the scheduler is optimal when a given DSR-CMOO problem is considered.

The idea is to find at each TTI *t* the optimal set of decision variables $\left\{c_{o,w_o}[t],u^{o}_{w_o,i}[t],b_{i,j}[t]\right\}$ in order to maximize the optimization problem $\left(P_{Primal}^{Agg}\right)$ and to respect the set of constraints $\left(C_{Primal}^{Agg}\right)$. Due to the product between the decision variables $\left(c_{o,w_o}[t]\cdot u^{o}_{w_o,i}[t]\cdot b_{i,j}[t]\right)$, the optimization problem $\left(P_{Primal}^{Agg}\right)$ becomes non-linear and thus, the optimal solution in not guaranteed. Based on

Equation 3.51, the multi-objective optimization problem can be divided into three categories when the dynamicity of the rule selection is taken into account:

1. **SSR based SMOO** problem $\left(\left(P_{Primal}^{Agg}\right)=\left(P_{Primal}^{Seq}\right)\right)$ if the objective $\forall o \in \mathcal{O}$ and the MUF $\forall w_o \in \mathcal{PU}_o$ are static over the entire transmission session. In this case, the SSR-SMOO problem refers to the classical optimization problems $\left(P_o\right)$, $\forall o \in \mathcal{O}$ analyzed in Sub-section 3.5.1 and constraints (a), (b), (c), (d) and (f) are not required.
2. **DSR based SMOO** problem $\left(\left(P_{Primal}^{Agg}\right)=\left(P_{Primal}^{Seq}\right)\right)$ if $\forall o \in \mathcal{O}$ is static over the entire downlink scheduling session and $w_o[t] \in \mathcal{PU}_o$ is variable TTI-by-TTI. Different MUFs are used concurrently in order to achieve the same target or objective, and constraints (f) consider only the satisfaction of the particular aggregate objective function. Chapter 6 presents the system model and simulation results when the DSR-SMOO problems are focusing on fairness-throughput tradeoff and GBR objective.
3. **DSR based CMOO** $\left(\left(P_{Primal}^{Agg}\right)=\left(P_{Primal}^{Conc}\right)\right)$ if both decision variables $o[t] \in \mathcal{O}$ and $w_o[t] \in \mathcal{PU}_o$ are variable at different TTIs. Different MUFs with different objective targets may be applied in order to achieve the aggregate objective concurrently. Only in this particular case, the aggregate multi-objective conditions or constraints (f) are fully taken into account. Chapter 7 shows the results when the DSR-CMOO problems are focusing on NGMN fairness, GBR, HoL delay and PDR objectives.

By applying the decision variable $c_{o,w_o}[t]$ to the aggregate programming problem of Eq. 3.51, the scheduler evolves from state $\mathcal{S}_t^S$ to $\mathcal{S}_{t+1}^S$. Due to the time dependence process, the optimization problem $\left(P_{Primal}^{Agg}\right)$ is *dynamic*. The newest state $\mathcal{S}_{t+1}^S$ contains the uncontrollable or the random subspace which does not depend on the applied decision variables in the previous TTI. For these reasons, Equation 3.51 is considered as a ***dynamic and stochastic optimization problem***.

It is clear that by optimizing such kind of combinatorial problems $\left(P_{Primal}^{Agg}\right)$ it is not a trivial job due to the fact that finding the best variables $c_{o,w_o}[t], u_{w_o,i}^{o}[t]$ and $b_{i,j}[t]$ at each TTI t requires a long way of searching for optimization. If one scheduling rule can be selected at the beginning of each TTI, $\left(P_{Primal}^{Agg}\right)$ becomes a simple optimization problem where those users with the maximum scheduling metrics are allocated to different resource blocks. Obviously, it will be very time consuming if the decisions on the scheduling rules are to be made at each TTI. Therefore, developing a policy of scheduling rule selection is one of the best ways to ease the decision making at each TTI.

The scheduling rule selection policy $\pi_{\nabla U}[t] = \left\{c_{o,w_o}[t]\right\}_{\forall o,w_o}$ represents a generic set of scheduling rules or marginal utilities that are applied dynamically TTI-by-TTI based on different circumstances from the scheduler state space. For example, a policy of marginal utilities can be decided as indicated in Eq. 3.52:

$$\pi_{\nabla U} = \left\{c_{1,3}[t], c_{3,2}[t+1], c_{2,5}[t+2], c_{4,3}[t+3]....\right\} \tag{3.52}$$

Therefore, reaching the optimal policies $\pi_{\nabla U}^{*}$ which establish the most representative scheduling rule at each TTI becomes the main focus of this research. The optimization and refinement of such sequences abovementioned are not a trivial job mainly because of the stochastic nature of the process which requires an infinite state space for searching the optimal solution. Two main approaches can be proposed for the policy optimization:

1. **Evolutionary methods**: e.g., expression and evolutionary programming;
2. **Dynamic programming methodologies**: e.g., real-time dynamic programming and temporal difference based learning algorithms such as ***reinforcement learning*** techniques.

Under the assumptions of constant power allocation and the sub-optimal MCS allocation, the system complexity of $\left(P_{Primal}^{Agg}\right)$ is $\mathcal{C}\left(|\mathcal{O}|\times|\mathcal{PU}_o|\times|\mathcal{U}_t|\times|\mathcal{B}|\right)$. Each algorithm above-mentioned requires a reasonable number of scenarios in order to fine tune the final policy for the real time downlink scheduling.

3.6.1 DSR based SMOO/CMOO Problems

As mentioned earlier, the aggregate MOO problem presented in Eq. 3.51 is a non-linear programming problem due to the product $c_{o,w_o}[t]\cdot u^{o}_{w_o,i}[t]\cdot b_{i,j}[t]$ between the scheduling rule decision variable, MU allocation variable and the resource allocation variable. Once the scheduling rule variable $c_{o,w_o}[t]$ is decided, the entire aggregate problem is reduced to a simple resource allocation procedure. There are proposed three ways in solving such kind of optimization problems:

1. **Sequential Problem Linearization**: converts the non-linear problem into its corresponding linear representation;
2. **Parallel Problem Linearization**: divides the non-linear MOO problem into $|\mathcal{PU}|$ linear sub-problems.
3. **Sequential Problem Linearization in Two Stages**: divides the non-linear MOO problem into two different stages of linear optimization problems.

All these linearization techniques constitute contributions of this research and each of these principles is analyzed in the following sub-sections.

3.6.1.1 Sequential Linearization

The MOO problem can be transformed into a linear optimization problem by introducing an additional variable $d_{o,w_o,i,j}[t]=c_{o,w_o}[t]\cdot u^{o}_{w_o,i}[t]\cdot b_{i,j}[t]$. The same principle of linearization is presented in [64] for the joint resource and MCS allocations. Then, the set of matrices which encompasses the rule assignment variables, the MU assignment variables and the resource allocation variables is: $d=\left\{d_{o,w_o,i,j}[t], o=1,..,|\mathcal{O}|, w_o=1,..,|\mathcal{PU}_o|, i=1,..,|\mathcal{U}_t|, j=1,..,|\mathcal{B}|\right\}$, where the entire set size is $|\mathcal{O}|\times|\mathcal{PU}_o|\times|\mathcal{U}_t|\times|\mathcal{B}|$. Then, the MOO problem becomes $\left(P^{Agg1}_{Primal}\right)$ and it is expressed in Eq. 3.53, where $l\in\mathbb{R}_+$ is a large positive number. Under this form, the global maximum solution is guaranteed since the overall optimization problem is a type of linear programming problem (the aggregate optimization is a sum of concave functions and the set of constraints $\left(C^{Agg1}_{Primal}\right)$ is convex).

$$\left(P_{Primal}^{Agg1}\right): \max_{\pi_{RB},\pi_{\nabla U}} \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|} d_{o,w_o,i,j}[t]\cdot U'^{o}_{w_o,i}\left(\mathcal{X}^{o}_{w_o,i}[t]\right)\cdot r_{i,j}[t]$$

$$\left.\begin{aligned} & d_{o,w_o,i,j}[t] \le b_{i,j}[t] \\ & d_{o,w_o,i,j}[t] \le c_{o,w_o}[t]\cdot l \\ & d_{o,w_o,i,j}[t] \le u^{o}_{w_o,i}[t]\cdot l \\ & d_{o,w_o,i,j}[t] \ge c_{o,w_o}[t]-\left(1-u^{o}_{w_o,i}[t]\right)\cdot l-\left(1-b_{i,j}[t]\right)\cdot l \\ & \quad \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t,\forall j\in\mathcal{B} \end{aligned}\right\} \quad (a)$$

$$\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} c_{o,w_o}[t]=1 \quad (b)$$

$$\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} u^{o}_{w_o,i}[t]=1,\ i=1,...,|\mathcal{U}_t| \quad (c)$$

$$\left(C_{Primal}^{Agg1}\right): \ s.t. \ \sum_{i=1}^{|\mathcal{U}_t|} u^{o}_{w_o^*,i}[t]=|\mathcal{U}_t|,\ w_o^*\in\mathcal{PU}_o, o\in\mathcal{O} \quad (d)$$

$$\sum_{i=1}^{|\mathcal{U}_t|} u^{o}_{w_o^{\otimes},i}[t]=0,\ w_o^{\otimes}=1,...,|\mathcal{PU}_o|, o=1,...,|\mathcal{O}|, \forall w_o^{\otimes}\neq w_o^* \quad (e)$$

$$\sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t]=1,\ j=1,...,|\mathcal{B}| \quad (f)$$

$$c_{o,w_o}[t]\cdot\Phi_{o,w_o}[t+1]\ge 0,\ o=1,...,|\mathcal{O}|, w_o=1,...,|\mathcal{PU}_o| \quad (g)$$

$$\begin{aligned} & d_{o,w_o,i,j}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t,\forall j\in\mathcal{B} \\ & c_{o,w_o}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o \\ & u^{o}_{w_o,i}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t \\ & b_{i,j}[t]\in\{0,1\},\ \forall i\in\mathcal{U}_t,\forall j\in\mathcal{B} \end{aligned} \quad (3.53)$$

The role of the first set of constraints (a) from $\left(C_{Primal}^{Agg1}\right)$ is to act as a *truth table* (AND gate), where the input variables $\left\{c_{o,w_o}[t];u^{o}_{w_o,i}[t];b_{i,j}[t]\right\}\in\{0,1\}$ count eight possible combinations of $\{0,1\}$ and parameter $d_{o,w_o,i,j}[t]\in\{0,1\}$ is the output variable. Based on this principle, when the output variable is $d_{o,w_o,i,j}[t]=1$, then all input variables must be equal to one such as $\left\{c_{o,w_o}[t];u^{o}_{w_o,i}[t];b_{i,j}[t]\right\}=\{1\}$ $\forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t,\forall j\in\mathcal{B}$. When the resource allocation is performed, only one objective, one MUF and one active user must be assigned for each RB. This involves the following constraint of the output variable as shown in Eq. 3.54:

$$\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|} d_{o,w_o,i,j}[t] = |\mathcal{B}| \tag{3.54}$$

Basically, Equation 3.54 is similar to $\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}\sum_{i=1}^{|\mathcal{U}_t|} d_{o,w_o,i,j}[t] = 1,\ j = 1,...,|\mathcal{B}|$. Due to the fact that the variable $d_{o,w_o,i,j}[t]$ replaces the product of three decision variables, the first set of constraints from Eq. 3.53 must be used instead of constraint from Eq. 3.54. Constraints (b) to (g) from Eq. 3.53 are similar with the set of constraints exposed by the initial optimization problem in Eq. 3.51. The reason behind the linearization procedure of $\left(P_{Primal}^{Agg1}\right)$ is to guarantee the global optimum solution when selecting the scheduling rule for a given scheduler state.

The linear MOO problem from Eq. 3.53 can be solved by using different integer programming approaches as suggested in [57], [58]. The computation complexity increases with the $|\mathcal{PU}|$ size, and this approach becomes immediately unsuitable for real-time LTE scheduling. More recently, the meta-heuristic methods such as genetic algorithms (GA), simulated annealing and Tabu-search are used to solve stochastic and dynamic combinatorial optimization problems [59], [60]. But all of these approaches proved a varied degree of success under discrete state space stochastic optimization [61]. The scheduler optimization problem is based on continuous and dynamic state space representation. Moreover, by increasing the $|\mathcal{PU}|$ size and the number of users, the optimization problem will suffer from the curse of dimensionality. Based on [62], [63], the meta-heuristics approaches show weak performance for large scale combinatorial problems in terms of the time complexity and the quality of results. By combining GA with other heuristics, the computation overhead becomes significant, which makes these approaches inappropriate for the real time LTE scheduling.

3.6.1.2 Parallel Linearization

In the optimization problems of non-linear $\left(P_{Primal}^{Agg}\right)$ and linear $\left(P_{Primal}^{Agg1}\right)$ formulations, the scheduling rule decision and MUF and RB assignments are jointly achieved. In order to reduce the computational complexity, the integer

non-linear problem $\left(P_{Primal}^{Agg}\right)$ can be divided into $|\mathcal{PU}|$ linear sub-problems which can be processed in parallel. Basically, this approach aims to run different LTE schedulers in parallel by performing different scheduling rules. After the assignment of RBs is performed for each parallel process, the scheduling rule which maximizes $\left(P_{Primal}^{Agg}\right)$ and respects the constraint set is selected. The family of scheduling rules focusing on the fairness performance requires an *infinite number of utilities* due to the parameterization of the PF rule. Then, this approach becomes infeasible when the NGMN fairness requirement is taken into account. More details about this concept are provided in Chapter 6.

3.6.1.3 Sequential Linearization in Two Stages

The main task of the linear optimization problem from Eq. 3.53 is to determine the best decision variable $c_{o,w_o}[t]$ at each TTI *t* in order to maximize the objective $\left(P_{Primal}^{Agg1}\right)$ and to respect the set of constraints $\left(C_{Primal}^{Agg1}\right)$. But this procedure does not guarantee the satisfaction of constraints (g) on Equation 3.53 which implicitly highlights the performance of the entire scheduling procedure when one scheduling rule has been applied. In order to tackle this problematic issue, the set of constraints (g) has to be included in the optimization problem of $\left(P_{Primal}^{Agg1}\right)$ by using relaxation methods. In this sense, the Augmented Lagrangian Function (ALF) and the dual optimization problem are required [90].

Let us define the Augmented Lagrangian function for the particular problem of Eq. 3.53 such as $\mathcal{L}_A\left(d,c,\Phi^{\mathsf{A}}\right)$ which is defined as follows:

$$\begin{aligned}
&\mathcal{L}_A : \mathbb{R}^{|\mathcal{O}|\times|\mathcal{PU}_o|\times|\mathcal{U}_t|\times|\mathcal{B}|} \times \mathbb{R}^{|\mathcal{O}|\times|\mathcal{PU}_o|} \times \mathbb{R}^{|\mathcal{O}|\times|\mathcal{PU}_o|} \rightarrow \mathbb{R} \\
&\mathcal{L}_A\left(d,c,\Phi^{\mathsf{A}}\right) = \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|} d_{o,w_o,i,j}[t]\cdot U'^{o}_{w_o,i}\left(\mathcal{X}^{o}_{w_o,i}[t]\right)\cdot r_{i,j}[t] + \quad (1) \\
&\qquad \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} \Phi^{\mathsf{A}}_{o,w_o}[t]\cdot c_{o,w_o}[t]\cdot\Phi_{o,w_o}[t+1] + \quad (2) \\
&\qquad \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} \frac{\mu_{o,w_o}}{2}\cdot\left(c_{o,w_o}[t]\cdot\Phi_{o,w_o}[t+1]\right)^2 \quad (3)
\end{aligned} \tag{3.55}$$

where the first term of Eq. 3.55 is the function to be maximized from the problem $\left(P_{Primal}^{Agg1}\right)$, the second term represents the Lagrange relaxation function and finally, the third one is the penalty function [90]. Basically, the ALF is considered to be a combination of Lagrange relaxation and penalty methods in solving complex constrained optimization problems [90]. In Equation 3.55, μ_{o,w_o} is the penalty factor and $\Phi_{o,w_o}^{\mathsf{A}}[t]$ is the accumulated Lagrange multiplier that has to be updated TTI-by-TTI. According to [90], for the Augmented Lagrangian approach, the Lagrange multiplier is updated by using the following formula:

$$\Phi_{o,w_o}^{\mathsf{A}}[t+1] = \Phi_{o,w_o}^{\mathsf{A}}[t] + \mu_{o,w_o} \cdot c_{o,w_o}[t] \cdot \Phi_{o,w_o}[t+1] \tag{3.56}$$

where $\Phi^{\mathsf{A}}[t] = \left\{\Phi_{o,w_o}^{\mathsf{A}}[t]\right\}$ is the matrix of Lagrange multipliers at TTI t and $\mu = \left\{\mu_{o,w_o}\right\}$ is the penalty matrix for each objective $o \in \mathcal{O}$ and for each marginal utility function $w_o \in \mathcal{PU}_o$.

Based on the defined matrix of Lagrange multipliers, let us define the concave Lagrange dual function $\mathcal{G}_A\left(\Phi^{\mathsf{A}}\right)$ which is defined as shown in Eq. 3.57:

$$\mathcal{G}_A : \mathbb{R}^{|\mathcal{O}|\times|\mathcal{PU}_o|} \to \mathbb{R}, \quad \mathcal{G}_A\left(\Phi^{\mathsf{A}}\right) = \sup_{d,c} \mathcal{L}_A\left(d, c, \Phi^{\mathsf{A}}\right) \tag{3.57}$$

The objective is to find the optimal Lagrange dual function $\mathcal{G}_A\left(\Phi^{\mathsf{A}*}\right)$ at each TTI t in such a way that:

$$\mathcal{G}_A\left(\Phi^{\mathsf{A}*}[t]\right) = \sup_{d,c}\left[\mathcal{L}_A\left(d[t], c[t], \Phi^{\mathsf{A}*}[t]\right)\right] \geq \mathcal{L}_A\left(d^*[t], c^*[t], \Phi^{\mathsf{A}}[t]\right) \tag{3.58}$$

where $d^*[t]$ and $c^*[t]$ are the optimal assignment matrices at TTI t and $\Phi^{\mathsf{A}*}[t]$ represents the optimal matrix of Lagrange multipliers being calculated online at each TTI t. In other words, the role of the Lagrange dual function is to learn the optimal Lagrange multipliers and to take the assignment decisions based on their optimized values at each TTI. When the learned matrix of Lagrange multipliers is optimal, then the scheduling decision variables are optimal (Eq. 3.58). The remaining task is to maximize the aggregate multi-objective function at TTI t+1. Based on these aspects, the dual optimization problem is highlighted in Eq. 3.59:

$$\left(P_{Dual1}^{Agg1}\right):\ \underset{\pi_{RB},\pi_{\nabla U}}{max}\left\{\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}\Phi_{o,w_o}^{\mathsf{A}}[t]\cdot c_{o,w_o}[t]\cdot\Phi_{o,w_o}[t+1]+\right. \quad (1)$$

$$\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}\frac{\mu_{o,w_o}}{2}\cdot\left(c_{o,w_o}[t]\cdot\Phi_{o,w_o}[t+1]\right)^2+ \quad (2)$$

$$\left.\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}d_{o,w_o,i,j}[t]\cdot U'^{o}_{w_o,i}\left(\mathcal{X}^{o}_{w_o,i}[t]\right)\cdot r_{i,j}[t]\right\} \quad (3)$$

$$\left(C_{Dual1}^{Agg1}\right):\ s.t.\quad \left.\begin{aligned} &d_{o,w_o,i,j}[t]\le b_{i,j}[t]\\ &d_{o,w_o,i,j}[t]\le c_{o,w_o}[t]\cdot l\\ &d_{o,w_o,i,j}[t]\le u^{o}_{w_o,i}[t]\cdot l\\ &d_{o,w_o,i,j}[t]\ge c_{o,w_o}[t]-\left(1-u^{o}_{w_o,i}[t]\right)\cdot l-\left(1-b_{i,j}[t]\right)\cdot l\\ &\qquad\forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t,\forall j\in\mathcal{B}\end{aligned}\right\} \quad (a)$$

$$\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}c_{o,w_o}[t]=1 \quad (b)$$

$$\sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|}u^{o}_{w_o,i}[t]=1,\ i=1,...,|\mathcal{U}_t| \quad (c)$$

$$\sum_{i=1}^{|\mathcal{U}_t|}u^{o}_{w_o^*,i}[t]=|\mathcal{U}_t|,\ w_o^*\in\mathcal{PU}_o,o\in\mathcal{O} \quad (d)$$

$$\sum_{i=1}^{|\mathcal{U}_t|}u^{o}_{w_o^{\otimes},i}[t]=0,\ w_o^{\otimes}=1,...,|\mathcal{PU}_o|,o=1,...,|\mathcal{O}|,\forall w_o^{\otimes}\neq w_o^* \quad (e)$$

$$\sum_{i=1}^{|\mathcal{U}_t|}b_{i,j}[t]=1,\ j=1,...,|\mathcal{B}| \quad (f)$$

$$\begin{aligned} &d_{o,w_o,i,j}[t]\in\{0,1\},\forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t,\forall j\in\mathcal{B}\\ &c_{o,w_o}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o\\ &u^{o}_{w_o,i}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall i\in\mathcal{U}_t\\ &b_{i,j}[t]\in\{0,1\},\ \forall i\in\mathcal{U}_t,\forall j\in\mathcal{B}\end{aligned} \quad (3.59)$$

The MOO problem expressed in Eq. 3.59 is a non-linear programming problem where the first two terms in the optimization problem aim to select the best scheduling decision matrix in order to maximize the accumulated Lagrange multiplier and the aggregate multi-objective function at TTI *t*+1, whereas the third term is the typical radio resource allocation procedure performed under the selected MU function. It is important to notice that by selecting the optimal matrix $c^*[t]$ in the first term, the second term is also maximized. Therefore, the

proposed sequential linearization method aims to split the non-linear optimization problem $\left(P_{Dual1}^{Agg1}\right)$ into *two sub-optimal linear sub-problems* by following the reasoning which is expressed below:

a) In the first stage, the scheduling rule $\left(c_{o,w_o}[t]\right)$ which maximizes the product between the accumulated Lagrange multiplier at TTI *t* and the aggregate multi-objective function at TTI *t*+1 must be selected (see Eq. 3.60.a); when the accumulated Lagrange multiplier is optimal for each given scheduler state and for each discipline, then the scheduling rule which maximizes the Lagrange multiplier for a given objective $o \in \mathcal{O}$ and MU function $w_o \in \mathcal{PU}_o$ is selected to be applied. Therefore, the selected scheduling rule maximizes the aggregate multi-objective function calculated for the entire set of objectives and active users at TTI *t*+1;

b) In the second stage, the allocation procedure of RBs for the active UEs is performed based on the selected rule from the first stage (Eq. 3.60.b).

$$\begin{aligned}
&\left(P_{Dual1}^{'Agg1}\right):\ \max_{\pi_{\nabla U}} \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} \Phi_{o,w_o}^{\mathsf{A}}[t]\cdot c_{o,w_o}[t]\cdot \Phi_{o,w_o}[t+1] \\
&\left(C_{Dual1}^{'Agg1}\right): s.t.\ \ \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} c_{o,w_o}[t] = 1 \\
&\qquad\qquad c_{o,w_o}[t] \in \{0,1\},\ \forall o \in \mathcal{O}, \forall w_o \in \mathcal{PU}_o
\end{aligned} \tag{3.60.a}$$

$$\begin{aligned}
&\left(P_{Dual1}^{''Agg1}\right):\ \max_{\pi_{RB}} \sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|} b_{i,j}[t]\cdot U'^{o}_{w_o,i}\left(\mathcal{X}^{o}_{w_o,i}[t]\right)\cdot r_{i,j}[t] \\
&\left(C_{Dual1}^{''Agg1}\right): s.t.\ \ \sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t] = 1,\ \ j = 1,...,|\mathcal{B}| \\
&\qquad\qquad b_{i,j}[t] \in \{0,1\},\ \forall i \in \mathcal{U}_t, \forall j \in \mathcal{B}
\end{aligned} \tag{3.60.b}$$

The only way to solve the linear optimization problem $\left(P_{Dual1}^{'Agg1}\right)$ is to select the decision variable $c_{o,w_o}[t]$ based on the accumulated Lagrange multiplier $\Phi_{o,w_o}^{\mathsf{A}}[t]$ in order to maximize the instantaneous aggregate multi-objective function $\Phi_{o,w_o}[t+1]$ for objective $o \in \mathcal{O}$ and marginal utility function $w_o \in \mathcal{PU}_o$ at TTI *t*+1. There are two main problems in selecting the optimal decision

variable $c^{*}_{o,w_o}[t]$ at each TTI. In the first instance, the policy $\pi_{\nabla U}$ has to be refined and improved and the accumulated Lagrange multiplier must be updated by many times. The procedure starts with the premise that the original form of $\pi_{\nabla U}$ is sub-optimal. The role of the first problem $\left(P^{'Agg1}_{Dual1}\right)$ is to select at each TTI a proper decision variable $c_{o,w_o}[t]$, in order to refine and to optimize the original policy of scheduling rules $\pi_{\nabla U}$ by updating the accumulated Lagrange multipliers. When the Lagrange multipliers are optimized, the dual problems $\left(P^{'Agg1}_{Dual1}\right)$ and $\left(P^{''Agg1}_{Dual1}\right)$ should provide optimal solutions when compared against the original problem $\left(P^{Agg1}_{Primal}\right)$. In the second instance, the refinement and the improvement of policies is practically impossible due to the lack of the scheduler state space when the Lagrange multiplier $\Phi^{\mathsf{A}}_{o,w_o}$ and the aggregate multi-objective function Φ_{o,w_o} are computed. To conclude, enhanced functions are required instead of using $\Phi^{\mathsf{A}}_{o,w_o}$ and Φ_{o,w_o} in order to obtain sustainable scheduling policies such as $\pi^{*}_{\nabla U}$. Even under these conditions, the task is not trivial since the behaviors of RRM environment and LTE packet scheduler under different states and different scheduling rules are totally unknown.

When performing the scheduling decision variable $c_{o,w_o}[t]$, the scheduler state evolves from $\mathcal{S}^{S}_{t}$ to $\mathcal{S}^{S}_{t+1}$. For simplicity, let us consider the scheduler state space to be discrete and $s[t]=\mathcal{S}^{S}_{t}$. Let us define the set of neighbor states $\mathcal{NS}^{S}_{t}$ from the scheduler state space where the scheduler state can evolve from TTI *t* to TTI *t*+1, and the next state is defined as $s[t+1]=\mathcal{S}^{S}_{t+1}\in\mathcal{NS}^{S}_{t}$. By using the terminology from the machine learning domain, the state-action function $Q_{o,w_o}\left(s[t]\right)$ and the reward $\mathcal{RW}^{o}_{s,w_o}[t+1]$ function are obtained based on the accumulated Lagrange multiplier and aggregate multi-objective functions:

$$\begin{aligned}&\Phi_{o,w_o}[t+1]\mapsto\mathcal{RW}^{o}_{s,w_o}[t+1],\quad \mathcal{RW}^{o}_{s,w_o}:|\mathcal{O}|\times|\mathcal{PU}_o|\times|\mathcal{NS}^{S}_{t}|\to\mathbb{R}\\&\Phi^{\mathsf{A}}_{o,w_o}[t]\mapsto Q_{o,w_o}\left(s[t]\right),\quad Q_{o,w_o}:|\mathcal{O}|\times|\mathcal{PU}_o|\times|\mathcal{S}^{S}_{t}|\to\mathbb{R}\end{aligned}\tag{3.61}$$

where the reward $\mathcal{RW}^{o}_{s,w_o}[t+1]$ is used to measure the scheduling performance of applying the decision variable $c_{o,w_o}[t]$ in the previous scheduler state $s[t]=\mathcal{S}^{S}_{t}$ and $Q_{o,w_o}(s[t])$ is the accumulated reward for the decision variable $c_{o,w_o}[t]$ being applied in the scheduler state $s[t]=\mathcal{S}^{S}_{t}$ for an infinite number of visits. Under the assumptions of discrete scheduler state space, $Q(s[t])=\{Q_{o,w_o}(s[t])\}$ is the matrix of accumulated rewards for the discrete scheduler state $\forall s[t]\in\mathcal{NS}^{S}_{t-1}$, where $o=1,...,|\mathcal{O}|$ and $w_o=1,...,|\mathcal{PU}_o|$. At the same time, $\mathcal{RW}^{o}=\{\mathcal{RW}^{o}_{s,w_o}\}$ is the instantaneous reward matrix for objective $o\in\mathcal{O}$, where $s=1,...,|\mathcal{NS}^{S}_{t}|$ and $w_o=1,...,|\mathcal{PU}_o|$. More details about these terminologies in continuous state and action spaces are provided in Chapter 5. By using the above notations, the dual optimization problem is highlighted in Eq. 3.62:

$$
\begin{aligned}
&\left(P^{'Agg1}_{Dual2}\right): \max_{\pi_{\nabla U}} \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} Q_{o,w_o}(s[t])\cdot c_{o,w_o}[t]\cdot \sum_{s=1}^{|\mathcal{NS}^{S}_{t}|}\cdot e^{o}_{w_o,s}[t]\cdot \mathcal{RW}^{o}_{s,w_o}[t+1] \\
&\left(C^{'Agg1}_{Dual2}\right): s.t. \quad \sum_{o=1}^{|\mathcal{O}|}\sum_{w_o=1}^{|\mathcal{PU}_o|} c_{o,w_o}[t]=1 \\
&\qquad\qquad \sum_{s=1}^{|\mathcal{NS}^{S}_{t}|} e^{o}_{w_o,s}[t]=1,\ o=1,...,|\mathcal{O}|, w_o=1,...,|\mathcal{PU}_o| \\
&\qquad\qquad c_{o,w_o}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o \\
&\qquad\qquad e^{o}_{w_o,s}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall w_o\in\mathcal{PU}_o,\forall s\in\mathcal{NS}^{S}_{t}
\end{aligned}
\tag{3.62}
$$

where $e^{o}_{w_o,s}[t]$ is the variable that decides the next state $s[t+1]$ and the future state assignation matrix becomes $e^{o}[t]=\{e^{o}_{w_o,s}[t], w_o=1,...,|\mathcal{PU}_o|, s=1,...,|\mathcal{NS}^{S}_{t}|\}$, $\forall o\in\mathcal{O}$. The optimization problem $\left(P^{'Agg1}_{Dual2}\right)$ is non-linear due to the product between the scheduling rule variable and the future state variable such as $c_{o,w_o}[t]\cdot e^{o}_{w_o,s}[t]$. In this sense, an additional variable $f_{o,w_o,s}[t]=c_{o,w_o}[t]\cdot e^{o}_{w_o,s}[t]$ is computed which follows the same principle of linearization which is exposed in Equation 3.53 and $f_o[t]=\{f_{o,w_o,s}[t], w_o=1,...,|\mathcal{PU}_o|, s=1,...,|\mathcal{NS}^{S}_{t}|\}$ is a matrix

which indicates the scheduler state evolution at TTI t+1 when the decision variable for the scheduling rule selection $c_{o,w_o}[t]$ has been applied in the previous state $\forall o \in \mathcal{O}, \forall w_o \in \mathcal{PU}_o$. The third form of the dual optimization problem is illustrated in Equation 3.63:

$$
\begin{aligned}
\left(P'^{Agg1}_{Dual3}\right): \max_{\pi_{\nabla U}} & \sum_{o=1}^{|\mathcal{O}|} \sum_{w_o=1}^{|\mathcal{PU}_o|} \sum_{s=1}^{|\mathcal{NS}_t^S|} Q_{o,w_o}\left(s[t]\right) \cdot f_{o,w_o,s}[t] \cdot \mathcal{RW}^o_{s,w_o}[t+1] \\
& f_{o,w_o,s}[t] \le c_{o,w_o}[t] \\
& f_{o,w_o,s}[t] \le e^o_{w_o,s}[t] \cdot l \\
& f_{o,w_o,s}[t] \ge c_{o,w_o}[t] - \left(1 - e^o_{w_o,s}[t]\right) \cdot l \\
& \quad \forall o \in \mathcal{O}, \forall w_o \in \mathcal{PU}_o, \forall s \in \mathcal{NS}_t^S \\
\left(C'^{Agg1}_{Dual3}\right): s.t. & \sum_{o=1}^{|\mathcal{O}|} \sum_{w_o=1}^{|\mathcal{PU}_o|} c_{o,w_o}[t] = 1 \\
& \sum_{s=1}^{|\mathcal{NS}_t^S|} e^o_{w_o,s}[t] = 1, \; o = 1,...,|\mathcal{O}|, w_o = 1,...,|\mathcal{PU}_o| \\
& c_{o,w_o}[t] \in \{0,1\}, \; \forall o \in \mathcal{O}, \forall w_o \in \mathcal{PU}_o \\
& e^o_{w_o,s}[t] \in \{0,1\}, \; \forall o \in \mathcal{O}, \forall w_o \in \mathcal{PU}_o, \forall s \in \mathcal{NS}_t^S \\
& f_{o,w_o,s}[t] \in \{0,1\}, \; \forall o \in \mathcal{O}, \forall w_o \in \mathcal{PU}_o, \forall s \in \mathcal{NS}_t^S
\end{aligned}
\tag{3.63}
$$

The idea is to select the best variable $c_{o,w_o}[t]$ which maximizes the accumulated reward value $Q_{o,w_o}\left(s[t]\right)$ and then, to assign the state $s[t+1]$ based on the assignation matrix $f_o[t]$ in order to maximize the instantaneous reward value at TTI t+1 $\mathcal{RW}^o_{s,w_o}[t+1]$. In real practice, when selecting the scheduling rule $c_{o,w_o}[t]$, the state $s[t+1]$ is assigned as a result of the scheduling procedure and the reward value is received from the RRM evaluation entity. The linear optimization problem exposed in Eq. 3.36 is theoretical rather than a practical one. The problem is to find the decision variable $c^*_{o,w_o}[t] = arg\,max\, Q^*_{o,w_o}\left(s[t]\right)$. Then, the scheduler reward at TTI t+1 is maximized when the optimal decision is applied. Similar to other control systems [42], the idea is to maximize the total expected return or the expected accumulated reward $\mathcal{RW}^A_{\pi_{\nabla U}}$ for a given policy of

rules $\pi_{\nabla U}$ starting from any initial state $s[t]$ until the optimal scheduler state is reached $\mathcal{S}_{t_{opt}}^{S} = s\left[t_{opt}\right]$. For optimality reasons, the accumulated reward for the considered policy is discounted according to Eq. 3.64:

$$\begin{aligned}\mathcal{RW}_{\pi_{\nabla U}}^{A}\left(s[t]\right) &= \mathcal{RW}_{\pi_{\nabla U}}\left(s[t+1]\right) + \ldots + \gamma^{T_{opt}} \cdot \mathcal{RW}_{\pi_{\nabla U}}\left(s\left[t+T_{opt}+1\right]\right) \\ &= \sum_{t_{opt}=1}^{T_{opt}} \gamma^{t_{opt}} \cdot \mathcal{RW}_{\pi_{\nabla U}}\left(s\left[t+T_{opt}+1\right]\right) \\ &= \mathcal{RW}_{\pi_{\nabla U}}\left(s[t+1]\right) + \gamma \cdot \mathcal{RW}_{\pi_{\nabla U}}^{A}\left(s[t+1]\right)\end{aligned} \tag{3.64}$$

where $\gamma \in [0,1]$ and $T_{opt} \to \infty$ due to the fact that the scheduling procedure is modeled as MDP problems with an infinite horizon. More details regarding to MDP modeling are provided in Chapter 5. The discount factor sets the importance of future rewards. Equation 3.65 is the case of *temporal difference learning* since:

$$\mathcal{RW}_{\pi_{\nabla U}}\left(s[t+1]\right) = \mathcal{RW}_{\pi_{\nabla U}}^{A}\left(s[t]\right) - \gamma \cdot \mathcal{RW}_{\pi_{\nabla U}}^{A}\left(s[t+1]\right) \tag{3.65}$$

When the instantaneous scheduler reward $\mathcal{RW}_{\pi_{\nabla U}}\left(s[t+1]\right)$ is equivalent with the difference between two accumulated rewards (Eq. 3.65) for a given policy $\pi_{\nabla U}$ for each possible scheduler states, then the considered policy $\pi_{\nabla U}^{*}$ is optimal. Basically, the accumulated reward value at state $s[t]$ for a given policy $\pi_{\nabla U}$ is similar to $Q_{o,w_o}\left(s[t]\right)$ where the decision variable $c_{o,w_o}[t]$ is an action being extracted from learned policy $\pi_{\nabla U}$. The policy has to be improved and evaluated for many visits of state $s[t]$ in order to learn the optimal scheduling rule $c_{o,w_o}^{*}[t]$ that maximizes the accumulated reward value $Q_{o,w_o}^{*}\left(s[t]\right)$. This stage is entitled the exploration stage since all possible scheduling rules have to be tested for a given scheduler state. After the scheduling policy is trained and becomes optimal, the learned policy can be tested in the exploitation stage. During the exploitation stage, the optimization problem $\left(P_{Dual3}^{'Agg1}\right)$ is satisfied since the scheduling rule that maximizes the accumulated reward $Q_{o,w_o}\left(s[t]\right)$ is selected and maximizes at the same time the reward value in the next state $\mathcal{RW}_{\pi_{\nabla U}^{*}}\left(s[t+1]\right)$.

3.6.1.4 RL in DSR-SMOO/CMOO Problems

As mentioned in the previous sections, the scheduler state space is continuous and multi-dimensional and practically, the size of neighbor states $\left|\mathcal{NS}_{i,t+1}^{S}\right| \to \infty$ for each scheduler state $\mathcal{S}_t^S$. Categorically, the simplest look-up table is not suitable in storing the previous learned values for all possible states. In this research, a novel architecture is proposed, which makes use of the MLPNN ***as a function approximation***. Then, the unvisited states are estimated based on other visited states. The role of MLPNN is to form a set of non-linear functions which can map each scheduler state in optimal decision variables c_{o,w_o}^{*}.

The TD learning is a prediction methodology which is used in this research in order to estimate the accumulated reward value $\mathbb{E}\left[\mathcal{RW}_{\pi_{\nabla U}}^{A}\left(\mathcal{S}_t^S\right)\right]$. In [42], the authors conclude that the TD learning is a combination of Monte Carlo (MC) and Dynamic Programming (D-P) methods. The MC method provides the expected return only at the end of the simulation session due to the non-episodic characteristic of the LTE scheduler. D-P can estimate the accumulated reward at each TTI, and moreover it can update the previous learned or the estimated value based on the current learned value. More precisely, in Eq. 3.65, at TTI t+1, $\mathbb{E}\left[\mathcal{RW}_{\pi_{\nabla U}}^{A}\left(\mathcal{S}_t^S\right)\right]$ is the MC target and is totally unknown. Even if the reward $\mathcal{RW}_{\pi_{\nabla U}}\left(\mathcal{S}_{t+1}^S\right)$ and the state $\mathcal{S}_{t+1}^S$ are known from the RRM and MOO entities, $\mathbb{E}\left[\mathcal{RW}_{\pi_{\nabla U}}^{A}\left(\mathcal{S}_{t+1}^S\right)\right]$ is an estimation value and represents the MC target as well. As shown in Eq. 3.65, the previous estimated value $\mathbb{E}\left[\mathcal{RW}_{\pi_{\nabla U}}^{A}\left(\mathcal{S}_t^S\right)\right]$ can be updated based on the new value of $\mathcal{RW}_{\pi_{\nabla U}}\left(\mathcal{S}_{t+1}^S\right)+\gamma\cdot\mathbb{E}\left[\mathcal{RW}_{\pi_{\nabla U}}^{A}\left(\mathcal{S}_{t+1}^S\right)\right]$ which represents the D-P property. For the particular case of RL, the error between the old learned value and the new one is used to update the previous learned value. It is important to notice that by updating the previous learned value based on the TD error, $\mathcal{RW}_{\pi_{\nabla U}}^{A}\left(\mathcal{S}_t^S\right)$ is decreased or increased by providing lower or higher probabilities for the decision variable $c_{o,w_o}[t]$ to be selected in order to transit the scheduler

state from $\mathcal{S}_t^S$ to $\mathcal{S}_{t+1}^S$. In the literature, this aspect is known as a policy refinement procedure. More details about the insights of RL algorithms in continuous state spaces and continuous action spaces are provided in Chapter 5.

Another problem is the scheduler state space dimension. By adding new objectives for the CMOO problem, the scheduler state space becomes larger. The number of radio bearers has a big impact on the scheduler state space. For the RL algorithm, this is an important issue because additional time steps for the learning procedure are required. So, the ***space aggregation procedure*** is absolutely necessary for the RL entity. Chapter 4 provides the necessary information on the aggregation procedure.

3.6.2 The Proposed Architecture for DSR based SMOO/CMOO Problems

From the architectural point of view, the multi-objective optimization based on the dynamic scheduling rule defines two modules: ***Marginal Utility State Informer*** **(MUSI)** and ***Marginal Utility Type Informer*** **(MUTI)**. The MUTI entity converts the action which is provided by the intelligent controller in the corresponding scheduling rule for the scheduler entity such as $\mathcal{A}_t^a \rightarrow c_{o,w_o}[t]$, where *a* is the index of the controller action. Based on the MUTI decision, MUSI provides the necessary parameters from the scheduler state in order to compute the corresponding scheduling metrics for each user and for each RB.

In Figure 3.2, the proposed architecture of the LTE scheduler is considered to be sub-optimal for the following reasons:

1. The MCS assignment procedure for the TB computation is omitted for the combinatorial optimization problem $\left(P_{Dual1}^{''Agg1}\right)$ and this is executed in a different stage. Even if the results provided in [64] indicate a degradation of the cell spectral efficiency of about 10% when the RB and MCS assignments are performed in separate stages, the computational complexity is much higher when the optimal scheme is used. From these

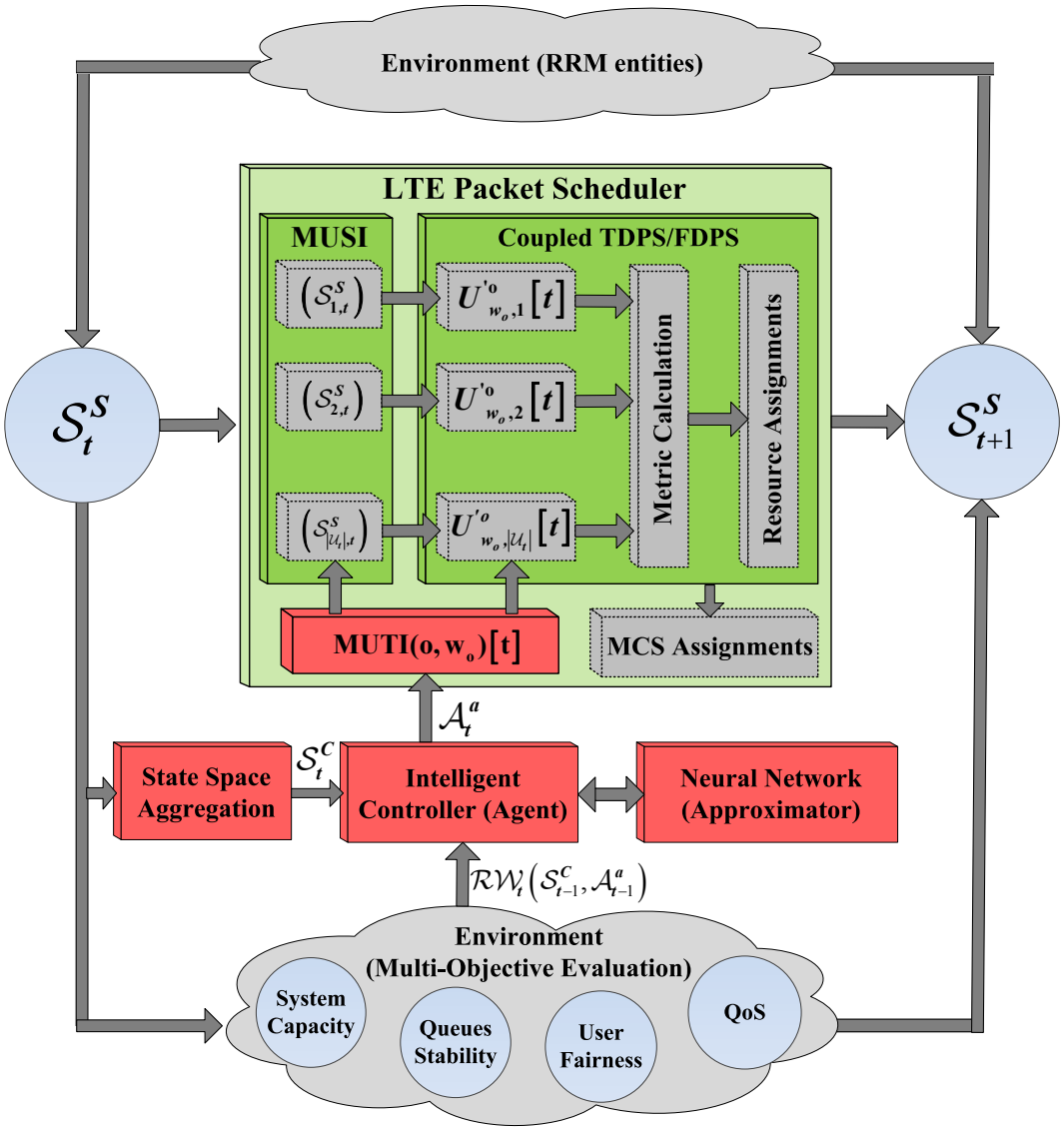


Fig. 3.2 The DSR based LTE Packet Scheduler Architecture

reasons, it is decided that in this approach, the MCS allocation for the selected RBs to be performed in a separate stage.

2. As shown in the previous sub-section, $\left(P_{Dual2}^{Agg1}\right)$ optimization is performed in two stages: $\left(P_{Dual2}^{'Agg1}\right)$ in which the best scheduling rule is learned based on various conditions, and $\left(P_{Dual2}^{''Agg1}\right)$ which performs the assignment of RBs. From this point, the scheduler is considered sub-optimal in the initial stage. Based on RL approaches, the intelligent agent is able to learn and to optimize the adopted policy $\pi_{\nabla U}$ until the optimality of the refined policy is reached ($\pi_{\nabla U}^{*}$). The policy optimization stage is also known as an

exploration stage. It is assumed that in the exploration stage, the scheduler is sub-optimal. The proposed *architecture exploits the optimized policy in what is called exploitation stage*. During the exploitation stage, *the learned scheduling policies are sustainable by maximizing the number of TTIs when the scheduler is considered optimal*. In practice, the sustainability of scheduling policies is tested in the exploitation stage by using the mean and STD values for percentage of feasible TTIs. Also, the reward type plays a crucial role since one objective of sustainable policies is to minimize the mean percentage of TTIs with punishment rewards.

The scheduler state space dimension depends on the system bandwidth and on the number of active users. From the controller perspective, the state space dimension should have a fixed dimension regardless of the parameters variation from the RRM entities. With an effective elimination of the irrelevant information from the scheduler state $\mathcal{S}_t^S$, the obtained set of parameters is considered to be the controller state space $\mathcal{S}_t^C$. The procedure is entitled the LTE ***scheduler state space aggregation***, and it is largely debated in Chapter 4. As mentioned earlier, even under the aggregate form of the scheduler state space, the obtained space is still continuous and the scheduling procedure remains non-episodic for some scenarios. Therefore, it is impossible first, to explore all the possible states, and second, to store the accumulated reward for all the visited states. For these reasons, the architecture makes use of MLPNN approximations by having the estimation role for the accumulated rewards. The MLPNN functionality aims to estimate the returns for the current and previous states and updates the learned value by training the MLPNN weights (details in Chapter 5).

The instantaneous reward value $\mathcal{RW}_{s,w_o}^{o}[t] = \mathcal{RW}_t\left(\mathcal{S}_{t-1}^S, \mathcal{A}_{t-1}^a\right)$ is provided based on the type of multi-objective optimization. Chapter 6 and Chapter 7 propose different reward functions based on DSR-SMOO/CMOO problems and present the sustainability of the proposed scheduling policies. The rest of the chapter is concentrated on the classification of scheduling techniques based on the achieved objectives. Following the proposed classification, the most relevant related work and studies are to be analyzed in terms of SMOO/CMOO problems.

3.7 The Proposed Classification of LTE Schedulers

The analyzed schedulers consider the CQI information in the utility computation in terms of achievable user rates for each RB and for each UE. For this reason, these techniques are entitled opportunistic schedulers whose ideas were introduced first in [65]. Unfortunately, some operators find this concept to be unpractical due to the higher complexity overhead introduced by the CQI feedback for each user and for each RB [66]. Consequently, *non-opportunistic* schedulers are still used regardless of the channel information, such as Round Robin (RR) for the best fairness in terms of number of allocated RBs, Earliest Deadline First (EDF), and Largest Weighted Delay First (LWDF) focusing on the packet delay, and alternatives for the service priorities such as Weighted Fair Queuing (WFQ) [67], [68]. Obviously, these metrics cannot be used for the $\left(P_{Dual3}^{'Agg1}\right)$ problem optimization due to the channel-unaware characteristics.

The OFDMA radio access interface is proposed in LTE in order to improve the system capacity based on the multi-user diversity principle and based on the channel aware schedulers. In order to reduce the complexity overhead, schedulers with *limited CQI feedbacks* are adopted instead of schedulers with *full CQI reporting schemes* [69], [70], [71], [72]. This approach attracts inevitably the performance degradation from the viewpoint of the total system throughput. For instance, in [69] the authors report a throughput loss of about 10-15% when a modified PF scheduler is used for a variety of limited CQI reporting schemes. Another method is to feedback the CQI reports in subsequent parts in different TTIs based on the user mobility as suggested in [70]. When only the RBs with the best CQI values are reported or the RBs with the CQIs included in a certain threshold from the maximum CQI value, the reported average throughput loss is about 4-8% for the FDPS scheduling as indicated in [71], [72]. It is clear that, each of these performances depends on certain circumstances without offering a guarantee that the average throughput loss is framed in a certain threshold for the general cases. This research considers, in the CQI state space aggregation, *a full and perfect CQI report without any error or delay in the reporting schemes* in order to collect as many as possible CQI reports. Practically, this case represents a

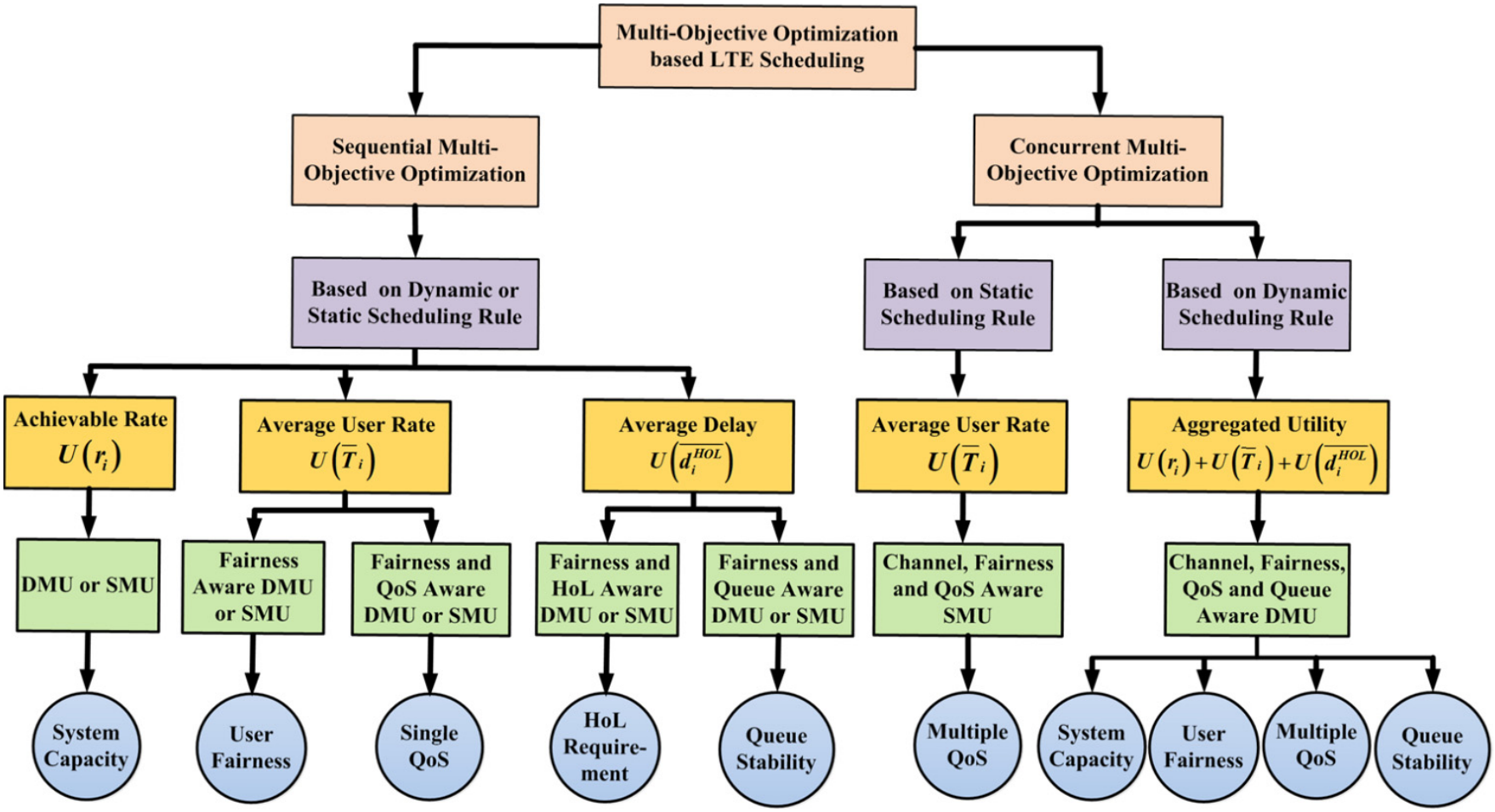


Fig. 3.3 The Classification of MOO based Opportunistic LTE Packet Schedulers

generalized approach and it can be applied very easily when other assumptions are considered in the CQI reporting process.

Alongside the above discussions, the LTE packet schedulers should consider the addressed MOO problem in their classification as indicated in Fig. 3.3. The SMOO based scheduling considers both cases of SSR and DSR approaches. Even in the case of DSR, the considered scheduling schemes are focusing only on one objective as discussed in the previous sections, adapting only the marginal utility function in order to refine the addressed policy. The DSR approach addresses the Dynamic Marginal Utility (DMU) whereas the SSR implies a Static Marginal Utility (SMU) over time.

The DSR-CMOO approach represents the main contribution of this research in which the sum of utility functions targeting multiple objectives is considered in the aggregate utility optimization problem. The MUF is changing TTI-by-TTI based on the RL approach targeting at each time instant different objectives. At the same time, the CMOO can be achieved by using a static scheduling rule. In this case, the MUF of the static scheduling rule can include multiple performance criteria such as HoL delay $d_i^{HoL}[t]$ and packet loss rate $R_i^{PL}[t]$ as indicated in [56], leading to the SSR-CMOO problems focusing on HoL delay and PLR. The only question that remains regarding the SSR based CMOO problem is the throughput optimality, issue which is unaddressed in [56].

3.8 Related Studies on MOO-Based Opportunistic LTE Scheduling

This section highlights the existing main results and contributions of the LTE scheduling techniques published in the literature under the classification scheme shown in Fig 3.3. The scheduling algorithms implemented in the previous radio access technologies which can be applied to OFDMA are also discussed for a more comprehensive understanding of the multi-objective optimization problem. The related work on the SMOO problems focusing on user fairness and system throughput tradeoff performance is analyzed in the first instance. Then the SMOO

approaches focusing on a single QoS objective are presented under the GBR and HoL packet delay criteria. The SMOO approaches focusing on system throughput maximization and queue stability are discussed in Appendix A.2 and Appendix A.6, respectively. The SSR-CMOO problems focusing on multiple targets are presented in Appendix A.7. Appendix A.8 resumes the state of the art in radio resource scheduling based on different DSR/SSR-SMOO/CMOO methodologies.

3.8.1 SMOO Focusing on User Fairness

By focusing only on finding system throughput optimal or near-optimal solutions will lead to an unfair treatment to users that experience a variety of CQI conditions. Hence, the fairness performance should be considered in advanced LTE networks in order to assure the minimum service level even for the elastic traffic types such as BE.

The problem of downlink LTE scheduling subject to adaptive fairness constraint is studied in [79]. The fairness performance is measured based on Jain Fairness Index (JFI), a metric which is determined by using the average user throughputs [80]. By adopting the JFI fairness requirement for the entire transmission session, the traditional PF scheduling rule is not able to respect the JFI constraint for different CQI state conditions. Therefore, in [79] a family of utilities based on the PF parameterization has been proposed in order to adapt the MUF based on the CQI state space conditions. The obtained scheduling rule is entitled the Generalized PF based on Simple Parameterization (GPF-SP), and the characteristics are highlighted by Eq. 3.66:

$$\begin{cases} U^{2}_{1(\alpha),i}\left(\overline{T}_i[t]\right)=\left(\overline{T}_i[t]\right)^{1-\alpha}\Big/\left(1-\alpha\right),\alpha\geq 0,\alpha\neq 1 \\ U^{2}_{1(\alpha),i}\left(\overline{T}_i[t]\right)=log\left(\overline{T}_i[t]\right),\alpha=1 \\ F'^{2}_{1(\alpha),i}\left(\overline{T}_i[t]\right)=1\Big/\left(\overline{T}_i[t]\right)^{\alpha} \\ D^{2}_{1(\alpha),i}\left(\overline{T}_i[t]\right)=r_{i,j}[t]\Big/\left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{3.66}$$

Basically, at each TTI, the best α parameter should be selected in order to maximize the optimization problem $\left(P_2\right)$ from Eq. 3.35 subject to JFI constraints.

The best PF parameter involves an infinite space of searching. Adopting sophisticated algorithms makes the approach unsuitable for the real time scheduling. It is the same case in [79], where the authors propose a novel technique which is able to predict the expected throughputs based on a probability mass function and then, the GPF parameter is changing based on the obtained JFI value. Unfortunately, this approach is required to be performed at the beginning of each scheduling decision, making it unsuitable for real time scheduling. The results suggest very good performance from the viewpoint of system throughput and user fairness tradeoff when compared with the original proposal such as the Second Order Cone Program (SOCP) [81]. The performance of this method is compared against the proposed scheduling policies in Chapter 6.

The JFI fairness requirement remains static for the whole transmission without being able to adapt to the novel conditions. In this sense, in [52] the fairness requirement based on CDF distribution has been introduced. The idea is to calculate the normalized user throughputs and to assure that at each TTI the CDF function should not cross the NGMN fairness requirement. It is in fact a dynamic requirement based on throughput observations which are changing at each TTI. When the CDF curve crosses the NGMN requirement, the system is declared unfair whereas staying too far from the requirement will push the system in the over-fairness area. Therefore, the feasible zone which is located in the neighborhood of the NGMN requirement on the right hand zone is proposed. In Chapter 6, the proposed scheduling policies show very good sustainability by maximizing the percentage of feasible TTIs when compared against the existing methodologies.

In [82] is proposed a novel architecture which adapts the MUF based on the NGMN requirement under a dynamic traffic load. The main idea is to implement a scheduler controller which performs the fairness adaptation at different time scales when compared with the LTE scheduler. The CDF fairness is analyzed in [82] based on traffic types with different rate requirements, but the GBR satisfaction condition is considered to be fulfilled and then, the fairness performance criterion becomes the main focus. Then, the scheduling rule based on the GBR requirement initially proposed in [83] is highlighted in Eq.3.67:

$$\begin{cases} U_{2,i}^{3}\left(\overline{T}_i[t]\right) = exp\left(\omega_{2,i}^{3} \cdot TC_i[t]\right) \cdot U_{1(\alpha),i}^{2}\left(\overline{T}_i[t]\right) \\ TC_i[t] = max\left\{0, TC_i[t-1] + \overline{\overline{T}}_i[t] - T_i[t]\right\} \\ F_{2,i}^{'3}\left(\overline{T}_i[t]\right) = 1 \Big/ \left(\overline{T}_i[t]\right)^{\alpha} \\ W_{2,i}^{3}\left(\overline{T}_i[t]\right) = exp\left(\omega_{2,i}^{3} \cdot TC_i[t]\right) \\ D_{2,i}^{3}\left(\overline{T}_i[t]\right) = exp\left(\omega_{2,i}^{3} \cdot TC_i[t]\right) \cdot r_{i,j}[t] \Big/ \left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{3.67}$$

The obtained scheduling rule is entitled GPF-SP with the Minimum/Maximum Rates (GPF-mM) where $TC_i[t]$ is the token counter, and the parameter $\omega_{2,i}^{3} > 0$ normalizes the MUF weight. When the number of active users is high, the Linear Mean Square Error (LMSE) approximation is used in order to calculate the minimum distance from the obtained CDF and the NGMN requirement. Then, α parameter is updated based on the minimum distance between LMSE-CDF and the NGMN requirement. Since the proposed scheduler adapts the fairness parameter at different time drops (multiple TTIs), the method becomes inflexible when aggressive changing in the traffic load may appear. The simulation results show that the indicated method is able to adapt the proposed MUF at each TTI in order to optimize the problem $\left(P_2\right)$ subject to NGMN fairness requirement with different GBR constraints and dynamic traffic loads. The proposal is considered to be coupled DSR-TDPS/FDPS based SMOO focusing on fairness since the GBR satisfaction is not addressed in [82].

Instead of adapting parameter α on the TTI basis [81], [82], the GPF parameterization can be achieved at each TTI per each RB [84]. Then, a new variable is needed as an argument for the utility function in terms of probabilities of allocating RB $j \in \mathcal{B}$ to UE $i \in \mathcal{U}_t$. The obtained argument is the expected throughput and the non-linear optimization problem is considered to be compact manifold where the extrema is found by using the Lagrange decomposition [84].

A generalized PF with Double Parameterization scheme (GPF-DP) for the user fairness and system throughput tradeoff control has been proposed in [85], [86]. Different tradeoff levels can be achieved by setting two parameters α and β as indicated by Eq. 3.68. It is shown in [86] that by using the coupled SSR-TDPS

$$\begin{cases} U^2_{2(\alpha,\beta),i}\left(\overline{T}_i[t]\right)=\left(r_{i,j}[t]\right)^{\beta-1}\cdot\left(\overline{T}_i[t]\right)^{1-\alpha}\Big/(1-\alpha),\alpha\in[0,1),\beta\in[0,1] \\ U^2_{2(\alpha,\beta),i}\left(\overline{T}_i[t]\right)=\left(r_{i,j}[t]\right)^{\beta-1}\cdot log\left(\overline{T}_i[t]\right),\alpha=1,\beta\in[0,1] \\ F'^2_{2(\alpha,\beta),i}\left(\overline{T}_i[t]\right)=1\Big/\left(\overline{T}_i[t]\right)^{\alpha} \\ W^2_{2(\alpha,\beta),i}\left(r_{i,j}[t]\right)=\left(r_{i,j}[t]\right)^{\beta-1} \\ D^2_{2(\alpha,\beta),i}\left(r_{i,j}[t]\right)=\left(r_{i,j}[t]\right)^{\beta}\Big/\left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{3.68}$$

/FDPS architecture for the (P_2) optimization problem from Eq. 3.35, a better throughput-fairness tradeoff can be achieved when compared with SSR-TDPS from the viewpoint of long-term and short-term JFI indicator.

Basically, for the coupled DSR-TDPS/FDPS, the double parameterization will have a much better impact in the optimization problem when the NGMN fairness requirement is considered under aggressive traffic load. By using the GPF-DP parameterization scheme based on the RL approach with continuous action state space, the obtained policies outperform any of other existing parameterization techniques when the NGMN requirement is considered in the reward function computation. By using the double parameterization policies, the percentage of feasible TTIs can increase with more than 35% when compared with the traditional approaches which consider the fairness parameterization. More details about this aspect are addressed in Chapter 6.

It is important to notice that the scheduling rule is PF, when $(\alpha=1,\beta=1)$, the rule is entitled MaxFair, when $(\alpha=1,\beta=0)$, and the discipline maximizes the system throughput (MT), when $(\alpha=0,\beta=1)$. In [87] is proposed a model for adaptive DSR-TDPS scheduling for UMTS standard subject to the fairness requirement when the case of $(\alpha=1,\beta=\text{var.})$ is considered in the optimization problem. The fairness requirement is defined in some acceptable limits. Parameter β is updated for each user in the long term purpose with the negative or positive steps. Such an approach outperforms the traditional PF from the viewpoint of user fairness performance when the SSR-TDPS scheduling technique is considered.

An option for the parameterized versions of PF introduced above is the Weighted PF (WPF) [88] in which the traditional GPF-SP or GPF-DP is replaced

by $m_j[t] = \underset{i \in \mathcal{U}_t}{arg\,max}\left\{ r_{i,j}[t] / \left(\overline{T}_i[t] / \Lambda_i[t] \right) \right\}$, where $\Lambda_i[t]$ is the fairness weight. The SMOO problem is non-convex, and the Lagrange dual decomposition may be used to find near-optimal solutions [89]. In [88], the fairness weights are determined based on the notion of utility fairness functions rather than on user rate fairness, as is the case of the previous methods. The obtained optimization problem is non-linear, and the interior point methods are used to obtained near-optimal solutions [90]. The simulation results are conducted through heterogeneous traffic types where the weights are adapted based on CQI reports and based on the traffic load. The proposed opportunistic fair method outperforms other algorithms such as WPF or MT from the fairness perspective with a very small degradation of the system throughput when compared with PF. The same WPF is studied in [91] where the weights are adapted in order to achieve the joint optimization problem of RBs, power and user assignments. Since such a problem is convex, it is proposed a distributed protocol which performs an online policy for the RB allocations, a heuristic algorithm for the power allocation, and a selfish strategy for the user assignments. The obtained results outperform the classical RR rule from the viewpoints of system throughput and user fairness strategies.

More details about the related work on SMOO problems focusing on throughput-fairness tradeoff performance are provided in Appendix A.3.

3.8.2 SMOO Focusing on GBR Objective

A critical task in LTE scheduling is to provide the GBR satisfaction in the sense that all active flows should respect the rate constraints from Table 2.1 based on different types of traffic.

The BF-PF scheduling rule presented in Sub-section 3.5.4 and evaluated originally for WCDMA scheduling [54], [99], can be further deployed for multi-carrier systems as shown in Eq. 3.41. However, the BF-PF scheme in HSDPA scheduling outperforms classical schemes such as PF and MT from the viewpoint of the percentage of satisfied streaming users [54]. For the rest of the thesis, it is preferable to use the notation of GPF-BF instead of BF-PF.

Two MUFs for PF and MT scheduling rules subject to GBR/MBR requirements for LTE systems are analyzed in [98] by using the parameterization of GPF-mM exposed in Eq. 3.67. This scheduling rule is originally proposed in [83], known as the Gradient algorithm with Minimum/Maximum Rate Constraints (GMR) under a token mechanism. It is shown in [83] that the GPF-mM rule is asymptotically optimal when $N_{TTI} \to \infty$. When applied to LTE systems, GMR-MT and GMR-PF outperform classical MT and persistent FDMA schedulers from the outage probability point of view [98]. The outage probability is defined here as a time fraction in which the average rate is below the GBR requirement.

The static GPF rule with the Required Activity Detection (GPF-RAD) subject to GBR constraints is studied in [94], [100] for TDPS scheduling and further implemented in OFDMA networks [101]. The GPF-RAD incorporates in the MUF weight the required scheduling rate. The utility function, the MUF and the scheduling rule for the GPF-RAD scheduling are shown in Eq. 3.69:

$$\begin{cases} U_{3,i}^{3}\left(\overline{T}_i[t]\right) = \breve{T}_i[t] \Big/ \overline{T}_i^{Sch}[t] \cdot U_{2(\alpha,\beta),i}^{2}\left(\overline{T}_i[t]\right) \\ \hat{\eta}_i[t] = \breve{T}_i[t] \Big/ \overline{T}_i^{Sch}[t] \\ F_{3,i}^{'3}\left(\overline{T}_i[t]\right) = 1 \Big/ \left(\overline{T}_i[t]\right)^{\alpha} \\ W_{3,i}^{3}\left(\overline{T}_i^{Sch}[t]\right) = \breve{T}_i[t] \Big/ \overline{T}_i^{Sch}[t] \cdot \left(r_{i,j}[t]\right)^{\beta-1} \\ D_{3,i}^{3}\left(\overline{T}_i^{Sch}[t]\right) = \hat{\eta}_i[t] \cdot \left(r_{i,j}[t]\right)^{\beta} \Big/ \left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{3.69}$$

where $\hat{\eta}_i[t]$ is the predicted required scheduling rate for UE *i* and $\overline{T}_i^{Sch}[t]$ is the scheduled average user rate and it is updated only and only if UE $i \in \mathcal{U}_t$ has been scheduled in the previous TTI [94]. In Chapter 6, the AUT-MMF observations are used instead of scheduled rate. It is important to notice that when $\sum_{i=1}^{|\mathcal{U}_t|} \hat{\eta}_i[t] \le 1$, the feasible load occurs with all users receiving the requested bit rate, and then, the second objective is the fairness performance. When $\sum_{i=1}^{|\mathcal{U}_t|} \hat{\eta}_i[t] > 1$, the congested case occurs and all users get the same degradation in the GBR satisfaction or the RAC module may decide to reject some users with lower priorities. Under heterogeneous traffic types with different GBR requirements, the

GPF-RAD outperforms GPF-BF and PF from the viewpoint of fraction of users who fails to achieve the GBR requirement as shown in [94] when the scheduling in HSDPA is performed. At the same time, GPF-RAD is able to schedule higher traffic load with heterogeneous GBR constraints, when compared against the static PF and GPF-BF $(\alpha = 1, \beta = 1)$ scheduling schemes [94].

The optimal scheduling under the guaranteed user rates gives rise to a problematic issue in OFDMA networks. The optimization problem of maximizing the sum of user rates under the rate constraints represents a non-linear mixed-integer optimization problem and is known to be NP-hard. Such problems are entitled *Generalized Assignment Problems* (GAP) [95]. Obviously, if one user is not selected for scheduling in the current TTI, its rate requirement cannot be respected. For this reason, setting a Time Window for data Rate Guarantee (TWRG) in which each user should guarantee its GBR/MBR becomes crucial. Under these concepts, a new problem arises: the balance between short-term/long-term GBR and the system throughput. If the TWRG is short, then only a part of active users can achieve their rate requirements in the short-term purpose, and the system throughput is degraded. If the time window is large, then the freedom degree is higher and the system throughput can be improved.

Based on the time window length which is used in the AUT-MMF observations, the NGMN fairness performance is strongly affected. In Chapter 6 are obtained different scheduling policies being focused on NGMN fairness criterion when the AUT-MMF observations are computed with different lengths of median moving filters. In Chapter 7, a novel RL approach is proposed being able to adapt the filter length in real time for the NGMN fairness, GBR and PDR objectives in order to increase the number of feasible TTIs.

In Chapter 6, different sustainable scheduling policies being oriented on the GBR objective are proposed by using different filter lengths when the AUT-MMF observations are computed for infinite buffer, CBR and VBR traffic types. A novel scheduling rule focusing on GBR constraint is proposed in Chapter 6 by using the Lagrange multiplier. When the infinite buffer traffic is scheduled, the proposed static scheme is the best option. More related work regarding the SMOO scheduling focusing on GBR requirement is presented in Appendix A.4.

3.8.3 SMOO Focusing on HoL Packet Delay Objective

The SMOO problems focusing on QoS parameters include the problem of delivering the packet to each UE in a given delay deadline. Several scheduling algorithms focusing on HoL delay have been studied in the literature. These algorithms will be presented in the following discussions. Based on the simple SSR schemes, more sophisticated schemes are discussed in Appendix A.5 based on decoupled TDPS/FDPS scheduling techniques.

The GPF-MLWDF scheme presented in Section 3.5.5 is probably the best known scheduling rule focusing on the delay requirement. Its predecessor LWDF rule optimality is shown in [107] and the M-LWDF for wireless systems was for the first time introduced in [55]. In [106] authors declare the M-LWDF unfair for those users who experience poor channel conditions when different packet discard timers are considered. A fairer M-LWDF scheme was proposed in this sense at a price of cell throughput degradation in the presence of non-elastic traffic types.

An alternative to GPF-MLWDF is the GPF based on the Exponential Function (GPF-EXP1) [108], [109] which is able to enhance the packet drop rate at the price of system throughput degradation when the streaming video service is used [109]. The utility function and the scheduling rule for GPF-EXP1 are expressed in Eq. 3.70. It is important to notice that for the BE traffic type, GPF-EXP1 acts as a pure PF scheduling rule.

$$\begin{cases} U_{2,i}^{4}\left(\overline{T}_i[t]\right) = exp\left[\left(\omega_{2,i}^{4}\cdot d_i^{HoL}[t] - \widehat{d^{HoL}}[t]\right)\Big/\left(1+\sqrt{\widehat{d^{HoL}}[t]}\right)\right]\cdot U_{2(\alpha,\beta),i}^{2}\left(\overline{T}_i[t]\right) \\ \widehat{d^{HoL}}[t] = 1/|\mathcal{U}_t|\sum_{i=1}^{|\mathcal{U}_t|}\omega_{2,i}^{4}\cdot d_i^{HoL}[t] \\ \omega_{2,i}^{4} = -log\left(\overline{R}_i^{PL}[t]\right)/\overline{d}_i^{HoL}[t] \\ F_{2,i}^{'4}\left(\overline{T}_i[t]\right) = 1\Big/\left(\overline{T}_i[t]\right)^{\alpha} \\ W_{2,i}^{4}\left(d_i^{HoL}[t]\right) = exp\left[\left(\omega_{2,i}^{4}\cdot d_i^{HoL}[t] - \widehat{d^{HoL}}[t]\right)\Big/\left(1+\sqrt{\widehat{d^{HoL}}[t]}\right)\right]\cdot\left(r_{i,j}[t]\right)^{\beta-1} \\ D_{2,i}^{4}\left(d_i^{HoL}[t]\right) = exp\left[\left(\omega_{2,i}^{4}\cdot d_i^{HoL}[t] - \widehat{d^{HoL}}[t]\right)\Big/\left(1+\sqrt{\widehat{d^{HoL}}[t]}\right)\right]\cdot\dfrac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}} \end{cases} \tag{3.70}$$

Two promising scheduling rules being oriented on the HoL delay are analyzed in [110], such as GPF-EXP2 and GPF-LOG scheduling disciplines when $(\alpha=1,\beta=1)$, as shown in Equation 3.71 and Equation 3.72 respectively. The set of parameters $\left(\omega_{3,i}^{4},\omega_{4,i}^{4}\right)$ is chosen according to [110]. The optimality of the GPF-EXP2 rule is discussed in [111], and based on the results from [110], GPF-EXP2 and GPF-LOG rules can support mixed QoS requirements by increasing at the same time the system capacity when compared with the GPF-MLWDF when the fairness parameters are: $(\alpha=1,\beta=1)$. The GPF-LOG throughput optimality is studied in [112] and it is shown that the GPF-LOG rule minimizes the overflow of data queues when compared with GPF-EXP2 and Maximum Weight rules [112].

$$\begin{cases} U_{3,i}^{4}\left(\overline{T}_i[t]\right)=exp\left[\left(\omega_{3,i}^{4}\cdot d_i^{HoL}[t]\right)\Big/\left(1+\sqrt{\widetilde{d_{ik}^{HoL}}[t]}\right)\right]\cdot U_{2(\alpha,\beta),i}^{2}\left(\overline{T}_i[t]\right) \\ \widetilde{d_{ik}^{HoL}}[t]=1/|\mathcal{U}_t|\cdot\sum_{ik=1}^{|\mathcal{U}_t|}d_{ik}^{HoL}[t],\ \forall ik\neq i \\ F_{3,i}^{'4}\left(\overline{T}_i[t]\right)=1\Big/\left(\overline{T}_i[t]\right)^{\alpha} \\ W_{3,i}^{4}\left(d_i^{HoL}[t]\right)=exp\left[\left(\omega_{3,i}^{4}\cdot d_i^{HoL}[t]\right)\Big/\left(1+\sqrt{\widetilde{d_{ik}^{HoL}}[t]}\right)\right]\cdot\left(r_{i,j}[t]\right)^{\beta-1} \\ D_{3,i}^{4}\left(d_i^{HoL}[t]\right)=exp\left[\left(\omega_{3,i}^{4}\cdot d_i^{HoL}[t]\right)\Big/\left(1+\sqrt{\widetilde{d_{ik}^{HoL}}[t]}\right)\right]\cdot\left(r_{i,j}[t]\right)^{\beta}\Big/\left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{3.71}$$

$$\begin{cases} U_{4,i}^{4}\left(\overline{T}_i[t]\right)=log\left(1+\omega_{4,i}^{4}\cdot d_i^{HoL}[t]\right)\cdot U_{2(\alpha,\beta),i}^{2}\left(\overline{T}_i[t]\right) \\ F_{4,i}^{'4}\left(\overline{T}_i[t]\right)=1\Big/\left(\overline{T}_i[t]\right)^{\alpha} \\ W_{4,i}^{4}\left(d_i^{HoL}[t]\right)=log\left(1+\omega_{4,i}^{4}\cdot d_i^{HoL}[t]\right)\cdot\left(r_{i,j}[t]\right)^{\beta-1} \\ D_{4,i}^{4}\left(d_i^{HoL}[t]\right)=log\left(1+\omega_{4,i}^{4}\cdot d_i^{HoL}[t]\right)\cdot\left(r_{i,j}[t]\right)^{\beta}\Big/\left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{3.72}$$

In [114] is proposed a modified version of classical Earliest Due to Date Function (EDF) [113] rule applied for LTE systems and entitled the GPF based on Modified EDF (GPF-MEDF) scheduling rule. The GPF-MEDF introduces a tunable function that can be set in order to satisfy heterogeneous delay requirements. The results conduct to the decision that GPF-MEDF is able to outperform GPF-MLWDF, GPF-LOG, GPF-EXP1, GPF-EXP2 $(\alpha=1,\beta=1)$ from the viewpoints of fairness, PLR and system throughput with a degradation of

the HoL packet delay when the number of video or VoIP flows increases. The GPF-EDF scheduling rule is expressed in Eq. 3.73.

$$\begin{cases} U_{5,i}^{4}\left(\overline{T}_i[t]\right)=1\Big/\left(\overline{d_i^{HoL}}[t]-d_i^{HoL}[t]\right)\cdot U_{2(\alpha,\beta),i}^{2}\left(\overline{T}_i[t]\right) \\ F_{5,i}^{'4}\left(\overline{T}_i[t]\right)=1\Big/\left(\overline{T}_i[t]\right)^{\alpha} \\ W_{5,i}^{4}\left(d_i^{HoL}[t]\right)=1\Big/\left(\overline{d_i^{HoL}}[t]-d_i^{HoL}[t]\right)\cdot\left(r_{i,j}[t]\right)^{\beta-1} \\ D_{5,i}^{4}\left(d_i^{HoL}[t]\right)=1\Big/\left(\overline{d_i^{HoL}}[t]-d_i^{HoL}[t]\right)\cdot\left(r_{i,j}[t]\right)^{\beta}\Big/\left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{3.73}$$

Another low complexity scheduling scheme with an exponential MU as a function of $d_i^{HoL}[t]$ and $\overline{d_i^{HoL}}[t]$ is proposed in [115]. The simulation results indicate that the proposed scheme increases the probability of transmitting packets in a given maximum-allowable delay for FTP, Web, VoIP and video services when compared against GPF-DP and GPF-MLWDF static rules when the considered set of fairness parameters is $(\alpha=1,\beta=1)$.

The reception buffer delay is considered in [116] as a component of the scheduler state space, and the results show a reduction of the play-out outage ratio when compared with GPF, GPF-MLWDF and GPF-EXP2 rules for $(\alpha=1,\beta=1)$, harming at the same time the system throughput performance. The results were conducted based on heterogeneous video and BE traffic types.

The SMOO scheduling problems focusing on the HoL delay presented so far, consider a single scheduling scheme for the mixed traffic. This approach is theoretical, rather than practical, since the scheduling scheme prioritizes the traffic classes. The traffic prioritization is a type of decoupled TDPS/FDPS scheduling since only a part of the total traffic load is passed to the FDPS module. More details about the HoL delay SMOO problems under the decoupled TDPS/FDPS scheduling architecture are described in Appendix A.5. Chapter 7 includes the scheduling rules presented in this sub-section and proposes the sustainable scheduling policies which can increase the percentage of feasible TTIs from the viewpoint of HoL delay and PDR requirements.

3.9 Summary

The sum rate maximization problem under QoS constraints is considered to be NP-hard. To avoid this drawback, different utilities are introduced in order to map the QoS objectives in the optimization problem. The results of this approach indicate linear programming models which aim to achieve different objectives by using the radio resource assignment matrices. The above approaches are known as the SSR-SMOO models since the entire optimization problem is balanced in the direction of a given scheduling objective based on the selected static marginal utility function. A mathematical model is developed for the DSR-SMOO/CMOO problems by targeting particular objective(s) when including different utility functions with different particularities in the aggregate optimization problem. The idea is to increase the time fraction or the number of TTIs when the scheduler is declared optimal from the viewpoint(s) of addressed objective(s). The obtained aggregate optimization problem is non-linear and the objective conditions from SSR-SMOO problems are considered constraints for the DSR-SMOO/CMOO proposal. In order to make the DSR-SMOO/CMOO problem tractable for the real time scheduling, the dual decomposition of the primal DSR-CMOO problem is performed. It was shown that by using the principles of Augmented Lagrangian method, the objective constraints are introduced in the optimization problem. For computational complexity reasons, sub-optimal schedulers are proposed by aiming to select in the first stage the scheduling rule and then, to perform the OFDMA resource assignments based on the selected scheduling rule in the first stage. However, the Augmented Lagrangian approach is not enough in finding sustainable policies of selecting scheduling rules due to the fact that the scheduler state space is not considered in the optimization problem. Therefore, the temporal difference learning methodology is introduced in the first dual optimization problem. The TD learning is used to estimate and to update the accumulated reward values for different states and scheduling rules in order to optimize a given policy. The reward function represents a discrete version of the aggregate objective functions averaged for each user. In order to make this approach suitable for real time scheduling, two approaches are imperiously required: the scheduler state space *approximation* and the scheduler *state space*

aggregation. Then, in the first optimization problem, the scheduling rule which maximizes the expected accumulated reward for a given state is selected. In order to learn the optimal value of the accumulated reward for a given state and for each rule, the policy improvement and evaluation techniques must be used by using different RL techniques which are introduced in Chapter 5. The idea is to learn based on the given state which is the most suitable scheduling rule to be applied in order to reach as fast as possible the feasible state for the considered DSR-SMOO/CMOO problems. The relevant studies and related work based on SMOO and CMOO optimizations show that the scheduling rules have different behaviors under different circumstances such as the traffic load, scheduler states and different assumptions. Under these considerations, different scheduling rules may be applied for the DSR-SMOO/CMOO problems in order to find the optimal scheduler states much faster with a reduced system complexity.

Chapter 4

LTE Scheduler State Space Aggregation

4.1 Chapter Outline

The LTE state space compaction is undoubtedly one of the most difficult tasks when the self-learning scheduling technique is used. On one side, the input state space dimension depends on the number of active radio bearers which practically makes the overall space usage impossible for the LTE scheduler controller. On the other side, a large input state space dimension requires more epochs of training in order to fine tune the sustainable policies of scheduling rules. Therefore, a very precise compaction modality of the original state space is imperiously needed. Due to the stochastic nature of the LTE scheduler, the general state space can be divided in two categories: *controllable state space* (which evolves based on an applied scheduling rule) and *uncontrollable state space* (regardless to the selected scheduling rule). The controllable elements refer to the multi-objective evaluation module and can be compacted by using statistical mathematical models. The uncontrollable elements require more sophisticated methods of compression. This chapter proposes a low complexity model for the LTE controllable and uncontrollable state space aggregation. For the CQI state space compaction, three offline stages are used in order to obtain high accuracy of the classified state space. The CQI preprocessing stage subtracts the overall CQI report state into a 15-dimensional state space. A novel procedure

of collecting the preprocessed CQI reports is proposed. Based on the collected data points, the unsupervised learning step is performed in order to obtain the optimal set of CQI data centers. The ***Simulated Annealing with Stochastic Tunneling*** (SAST) as a meta-heuristic method is proposed in order to avoid the local minima problems which exist in the traditional clustering approaches. The simulation results show that the proposed SAST method performs much better than the existing clustering methods, being able to minimize the average distortion between the obtained set of data centers and the preprocessed CQI data set. In order to classify the unobserved data sets which are not included in the collected data base, the additional supervised learning step is applied. In this sense, the ***SAST based Radial Basis Function Neural Network*** **(RBFNN)** ***with feed-forward and backward propagation*** is trained based on the obtained set of centers provided by the unsupervised learning step. The experimental results show that RBFNN is a very powerful tool in the CQI state space classification, minimizing the mean square error between the preprocessed input CQI state and the predicted output. The learning structure proposed in this chapter takes the advantage of the offline procedure. When the overall structure is trained, the classified output state space can be used under different forms of regressed values and applied directly to the input state space of the scheduler controller.

4.2 LTE Scheduler State Space Characteristics

As mentioned in Chapter 3, the overall scheduler state space evolution can be partially controlled by the scheduler decision. Basically, the list of parameters which is directly implied in the desired MOO problem (average user throughput, HoL packet delay, packet loss rate and queue size) is impacted by the scheduling decision variable $c_{o,w_o}[t]$, $\forall o \in \mathcal{O}$ and $\forall w_o \in \mathcal{PU}_o$. For these reasons, these parameters constitute the controllable LTE subspace. Unfortunately, other variables such as CQI reports, ACK/NACK RLC acknowledgements and packet arrival rate are not able to be adjusted by the selected scheduling rule. At the beginning of each TTI, the scheduler controller should be able to select proper scheduling rules for given sets of controlled and uncontrolled elements in order to

increase as much as possible the percentage of feasible/optimal TTIs and to improve the sustainability of the obtained scheduling policies in the long term purpose for a given DSR-SMOO/CMOO problem.

The direct dependency between the controllable and uncontrollable spaces makes the scheduling decision even more complicated. For instance, the uncontrollable parameters can be seen as instantaneous and stochastic variables whereas the controllable ones can be viewed as averaged parameters over time based on uncontrollable variables at the current time of scheduling. It is the case of the average throughput for UE $i \in \mathcal{U}_t$ at TTI *t* which is calculated based on the achieved instantaneous throughput (if UE $i \in \mathcal{U}_t$ was scheduled at TTI *t*-1) and based on the history of the average rate that depends on the type and on the length of the filter which is used. The evolution of the controllable parameters is based equally on the scheduling decision and on the instantaneous uncontrollable parameters. The mentioned concept can be expressed based on Eq. 4.1:

$$\mathcal{S}_i^{S,C}[t+1] = \mathcal{S}_i^{S,U}[t] \times \mathcal{R}[t] + \mathcal{S}_i^{S,C}[t] \tag{4.1}$$

where $\mathcal{S}_i^{S,C}$ and $\mathcal{S}_i^{S,U}$ represent the controllable and uncontrollable scheduler sub-spaces, respectively for each user $i \in \mathcal{U}_t$, and $\mathcal{R}[t] = \{c[t], u^o[t], b[t]\}$ is the decision set which can contain the resource allocation matrix $b[t]$, the objective and MU selection matrix $c[t]$ and the MU assignment matrix $u^o[t]$ for objective $\forall o \in \mathcal{O}$. In Equation 4.1, the unknown variable is denoted by the set of decision matrices $\mathcal{R}[t]$. The resource allocation variable $b_{i,j}[t]$ is determined based on the controllable and uncontrollable elements $\{\mathcal{S}^{S,U}[t], \mathcal{S}^{S,C}[t]\}$. The scheduler controller has to decide which scheduling rule should be applied at TTI *t*. Then, the scheduling decision can be viewed as a function expressed in Eq. 4.2:

$$\mathcal{R}[t] = \underset{o,w_o}{arg\,max}\left[\mathcal{F}_c\left(\mathcal{S}^{S,U}[t], \mathcal{S}^{S,C}[t]\right)\right] + \underset{i,j}{arg\,max}\left[\mathcal{F}_b\left(\mathcal{S}^{S,U}[t], \mathcal{S}^{S,C}[t]\right)\right] \tag{4.2}$$

where $\mathcal{F}_b(\cdot)$ and $\mathcal{F}_c(\cdot)$ are the resource allocation and scheduling rule functions to be maximized. Consequently, the LTE controller provides at each TTI an

eligible function $\mathcal{F}_c(\cdot)$ which is able to select proper scheduling rules in order to drive the evolution of the controllable state space in the feasible region requested by the MOO performance evaluation entity.

Due to the state space dimension sensitivity of the RL algorithms, the initial set of controllable and uncontrollable state spaces have to be transformed in a more compact representation in order to speed up the learning procedure and to converge to the optimal policy of scheduling rules. Based on the original LTE scheduler state space transformation, the obtained state space eliminates the number of users and the system bandwidth dependencies which are considered to be the redundant information for the LTE controller state space. For these reasons, the transformed scheduler state space is entitled the ***LTE controller state space***.

Based on the mentioned concept, the original scheduler state space should be transformed or ***aggregated*** into the controller state space, and the controller should learn the optimal function for each scheduling rule based on the compacted state space as shown by Eq. 4.3:

$$c[t] = \underset{o,w_o}{arg\,max}\, \mathcal{F}_c\left(\mathcal{S}^{C,U}[t], \mathcal{S}^{C,C}[t]\right) \tag{4.3}$$

where $\mathcal{S}^{C,U}[t], \mathcal{S}^{C,C}[t]$ are the uncontrollable and controllable aggregate state spaces, respectively. Both compacted state spaces reduce the original dimension to the most representative parameters of the original state space such as:

$$\begin{aligned} \mathcal{S}^{S,U}[t] = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{S}_i^{S,U}[t] &\xrightarrow[Functions]{Aggregation} \mathcal{S}^{C,U}[t] = \bigcup_{p_{CU}=1}^{N_{PCU}} \mathcal{S}_{p_{CU}}^{C,U}[t] \\ \mathcal{S}^{S,C}[t] = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{S}_i^{S,C}[t] &\xrightarrow[Functions]{Aggregation} \mathcal{S}^{C,C}[t] = \bigcup_{p_{CC}=1}^{N_{PCC}} \mathcal{S}_{p_{CC}}^{C,C}[t] \end{aligned} \tag{4.4}$$

where N_{pCU} and N_{pCC} are the number of uncontrollable and controllable aggregate parameters, respectively. From the functionality point of view, the scheduler state space compaction procedure precedes the scheduling rule selection and the RB allocations as depicted in Fig. 4.1. The aggregation functions for both scheduler spaces take different forms depending on the different type of information which is required by the scheduler controller. For the uncontrollable CQI reports, the controller should receive the general statement of the channel

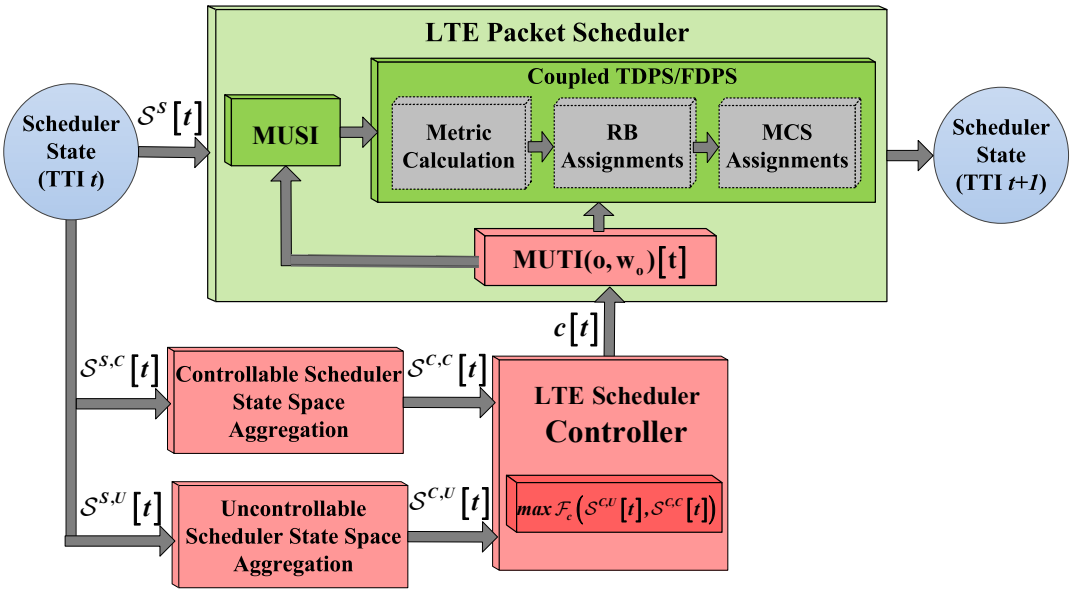


Fig. 4.1 The Aggregation of Uncontrollable and Controllable Scheduler State Spaces

qualities for all active users at each TTI. For the controllable parameters such as average user throughputs calculated with different types of filters, packet loss rates or HoL delays, the general statistics in terms of mean and STandard Deviation (STD) are enough to fine tune the optimal policy of scheduling rules. The details of the proposed aggregation functions for both controllable and uncontrollable scheduler spaces are discussed in the following sub-sections.

4.2.1 Controllable LTE Scheduler State Aggregation

The controllable set of parameters from the scheduler state space represents the most important indices for the MOO performance evaluation. Based on the proposed observations, the scheduler senses how far or close the overall system is from the imposed multi-objective target. From the controller decision point of view, the original state space engages two main drawbacks:

- The controllable state space dimension becomes very large when the numbers of users, objectives and priority classes increase.
- The controller observations depend on the number of users during the transmission session. This fact complicates the controller architecture since the state space dimension varies based on the number of active bearers which is undesirable for the RL training procedure.

Only the most relevant information should be passed onto the controller input state space that retains a general variation of the original observations reported to their average (mean) value. It is the case of the STD parameter which is intensively used in statistics and probabilistic systems [197]. By using the STD function for the controllable set of parameters, the controller is aware of how close the original set of observations is from its mean value. When the STD is low, the observation population is very close to the mean value whereas a large STD value implies a large spreading factor of the original set of observations. For the fairness performance evaluation, a lower STD of the average or mean user throughputs indicates that the scheduler is over-fair, and when the STD is very large, the system can be declared unfair. Therefore, the controller should be able to learn how to act in order to meet the optimum STD value in which the scheduler locates the fairness feasible zone. The same concept is applied for other DSR-SMOO/CMOO problems (to be detailed in Chapters 6 and 7).

The definition domain of the controller function which has to be approximated is $\mathcal{F}_c : \mathbb{R}_{[-1,1]} \rightarrow \mathbb{R}_{[-1,1]}$ that reveals the fact that both mean and STD values should be normalized. In practice, the upper and lower bounds of STD are very difficult to be reached since the controllable parameters take different forms of large real numbers. One way is to normalize at each TTI the original set of controllable observations based on their mean (expected) value.

For simplicity, let us consider the set of initial observations or the controllable set of parameters for each UE $i \in \mathcal{U}_t$ to be denoted by x_i^g, where $x_i^g[t] \in \left\{T_i, \overline{T}_i, d_i^{HoL}, \overline{d_i^{HoL}}, q_i^{TX}, \overline{q_i^{TX}}, R_i^{PL}, \overline{R_i^{PL}}\right\}$ and $g = 1,...,N_{par}^{SC}$, where N_{par}^{SC} is the number of scheduler controllable parameters. Then, the normalized controllable observation for UE $i \in \mathcal{U}_t$ can be expressed as follows:

$$\widehat{x_i^g}[t] = x_i^g[t] \Big/ \frac{1}{|\mathcal{U}_t|} \sum_{i=1}^{|\mathcal{U}_t|} x_i^g[t] \tag{4.5}$$

where the mean of the controllable normalized observations respects the property of $\left(1/|\mathcal{U}_t|\right) \cdot \sum_{i=1}^{|\mathcal{U}_t|} \widehat{x_i^g}[t] = 1$. Each normalized controllable observation can be seen

as a product of random variables such that $\widehat{x_i^g}[t] = \prod_{ii=1}^{|\mathcal{U}_t|} x_{ii}^g[t] / x_{ii+1}^g[t] \cdot \widehat{x_{ii+|\mathcal{U}_t|}^g}[t]$, $\forall ii \neq i \in \mathcal{U}_t$. Consequently, the normalized observation can be modeled as a lognormal variable and the corresponding STD and mean parameters can be determined by using the maximum likelihood estimation [197], [198]:

$$\mu_{\widehat{x_i^g}} = \frac{1}{|\mathcal{U}_t|} \sum_{i=1}^{|\mathcal{U}_t|} ln\left(\widehat{x_i^g}[t]\right) \tag{4.6}$$

$$\sigma_{\widehat{x_i^g}} = \sqrt{\frac{1}{|\mathcal{U}_t|} \sum_{i=1}^{|\mathcal{U}_t|} \left[ln\left(\widehat{x_i^g}[t]\right) - \mu_{\widehat{x_i^g}} \right]^2} \tag{4.7}$$

where $\mu_{\widehat{x_i^g}}$ and $\sigma_{\widehat{x_i^g}}$ are the mean and the standard deviation, respectively for the lognormal distributions of the normalized controllable scheduler parameters. The aforementioned representation for the controller input offers a very good representation for the original set of controllable parameters. For instance, when the mean value of user throughputs is $\mu_{\widehat{T_i}} = 0$, it means that the normalized averaged user throughputs are very close to 1 leading to a lower STD value and the scheduler can be considered over-fair. When the mean value is very close to $\mu_{\widehat{T_i}} \approx -1$, the normalized observations are located in the lower side of the mean value of 1, implying a higher standard deviation and the scheduler can be considered unfair. More details about the state space representation under the NGMN fairness criterion are provided in Chapter 6.

Correlated with the uncontrollable channel information that will be analyzed in the following sub-sections, the scheduler controller should be able to find the optimal mean and STD values $\mu_{\widehat{x_i^g}}, \sigma_{\widehat{x_i^g}}$ in which the scheduler operates in the desired feasible area imposed by the multi-objective optimization problem. The advantage of such representation refers to the fact that the definition domain of mean and STD $\mu_{\widehat{x_i^g}}, \sigma_{\widehat{x_i^g}}$ functions is much more reduced when compared with the traditional representation of the controllable parameter distribution. Additional parameters relative to the objective requirements are introduced in the controller state space, and the details will be analyzed in Chapters 6 and 7.

4.2.2 Uncontrollable LTE Scheduler State Aggregation

As mentioned in Sub-section 3.3.1 from Chapter 3, the uncontrollable scheduler space refers mainly to channel conditions, ACK/NACK notifications and packet arrival rates. For the last two observation types, the same concept described in the previous sub-section can be used as a form of aggregation when a single queue is considered for each radio bearer. In the case of multiple service rates, the scheduler priority mode should be activated, and the mean and STD parameters are calculated for each traffic priority type. However, one of the most important uncontrollable parameters is the CQI report. Since the controller should take scheduling decisions at each TTI based on the controlled and uncontrolled state spaces, the CQI report compaction becomes even more important.

The most common parameter which is used in the literature to describe the average radio conditions is the geometry factor (G-factor) which is calculated based on different PHY and radio condition parameters [40]. In terms of the scheduling decision, this parameter is not efficient since the scheduling is performed based on the quantized CQI reports. Therefore, the ***instantaneous general radio conditions*** should be obtained based on the received CQI reports. The rest of this chapter is concentrated on classifying the CQI user feedbacks in different clusters. Under this approach, the feedback scheme can be further simplified by reporting only the cluster index in which each user channel condition stands in.

4.3 Motivation for CQI State Space Aggregation

The radio channel introduces *frequency diversity* due to the multipath propagation [150],[151]. When users or obstacles are moving during transmission, different channel conditions improve or deteriorate. This concept is called *temporal diversity*. Moreover, because of the statistical independence of the user's fading processes, it is likely that by increasing the number of active users better channel quality feedbacks can be obtained. It is the case of the *multiuser diversity* that represents one way to increase the system capacity by using the opportunistic schedulers. In order to exploit the frequency, temporal and multiuser diversities,

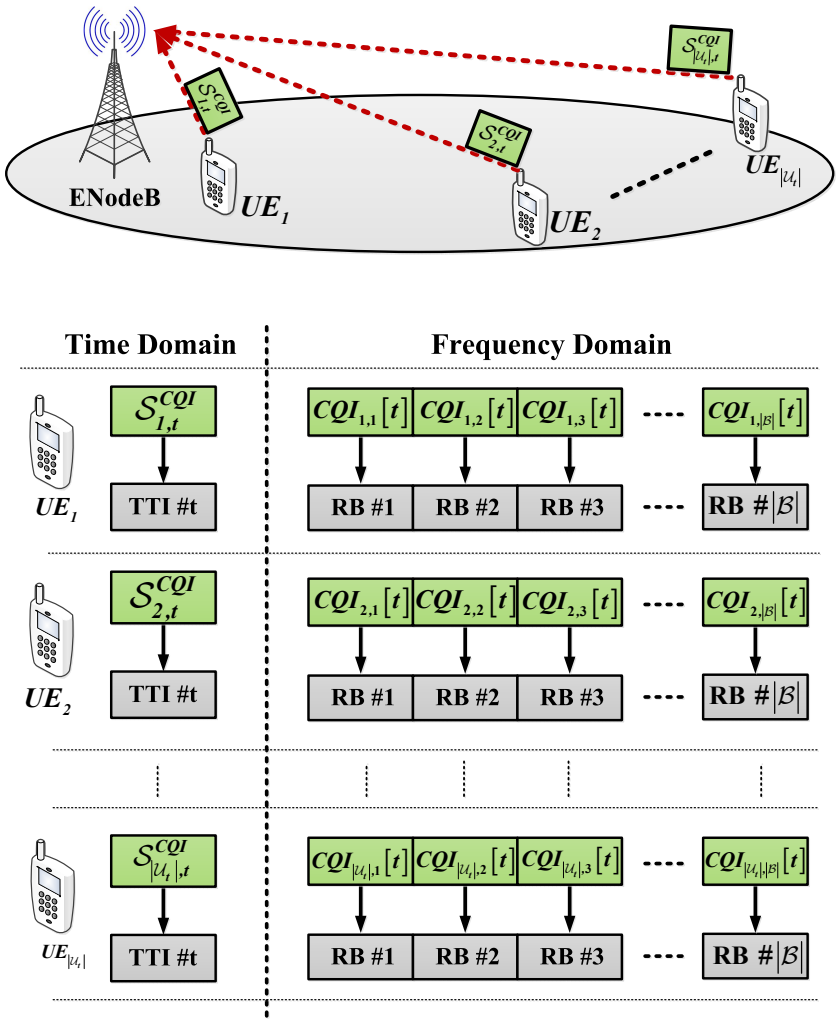


Fig. 4.2 Channel Quality Indicator Reports

the use of the CQI feedbacks becomes a mandatory and crucially task for the overall system performance (Fig. 4.2). If the LTE scheduler aims to increase the system performance balanced in the direction of the addressed objective, the LTE controller role is to increase the number of feasible TTIs when different DSR-SMOO/CMOO approaches are considered.

In the LTE standard, mobile terminals are configured to report to eNodeB through PUCCH channel different measurements of the downlink channel, such as CQI, Pre-coding Matrix Indicator (PMI) and Rank Indicator (RI). In general, PMI is relevant for the spatial multiplexing when the multiuser MIMO technology is used [149], [150]. The RI report is an indicator which informs how many independent data streams can be supported by different mobile users with the same MIMO-MU technique. If the power allocation, scheduling procedure and

MCS allocation are considered in the LTE scheduler functionality, the CQI reports can provide a prediction model of the general radio conditions to the LTE scheduler controller. Basically, there are two main types of reasons why the controller should consider the CQI state space compaction:

1. **Quantitative Reasons**: refer strictly to the CQI state $\mathcal{S}_t^{CQI}$ dimension depending on the number of active users and the traffic load fluctuations.
2. **Qualitative Reasons**: address the irrelevant CQI information and the *improvement of the opportunism loss effect*.

As mentioned in Chapter 3, the balance between system throughput and other QoS objectives is achieved based on the opportunistic schedulers when the CQI reports are considered in the scheduling metric computation. In order to take the advantage of the opportunistic scheduling for the LTE scheduler controller, the CQI aggregation requires a very high precision of the CQI statement in the aggregate controller state space $\mathcal{S}_t^C$.

4.4 The Proposed Architecture for the CQI State Space Aggregation

Based on quantitative and qualitative motivations, a novel CQI aggregation scheme is proposed in this section as an interaction module between the CQI state space and the LTE scheduler controller (Fig. 4.3). The avoidance of time-frequency dependency implies a ***preprocessing stage*** **(CQI-PS)** in which the number of elements of the CQI state space $\mathcal{S}_{i,t}^{CQI}$ for each user $i \in \mathcal{U}_t$ becomes constant regardless of the number of active users in the cell and regardless of the system bandwidth. The redundant data extraction is based on the CQI-PS and implies the ***classification stage*** **(CQI-CS)**. The classified CQI state space depends on the number of classes. If the number of classes is high, then the obtained state is significant for the controller. In this sense, the ***regression stage*** **(CQI-RS)** is required in order to eliminate the direct dependency on the number of preprocessed CQI classes. In the following, each stage is shortly explained and more comprehensive explanations are provided in the upcoming sub-sections.

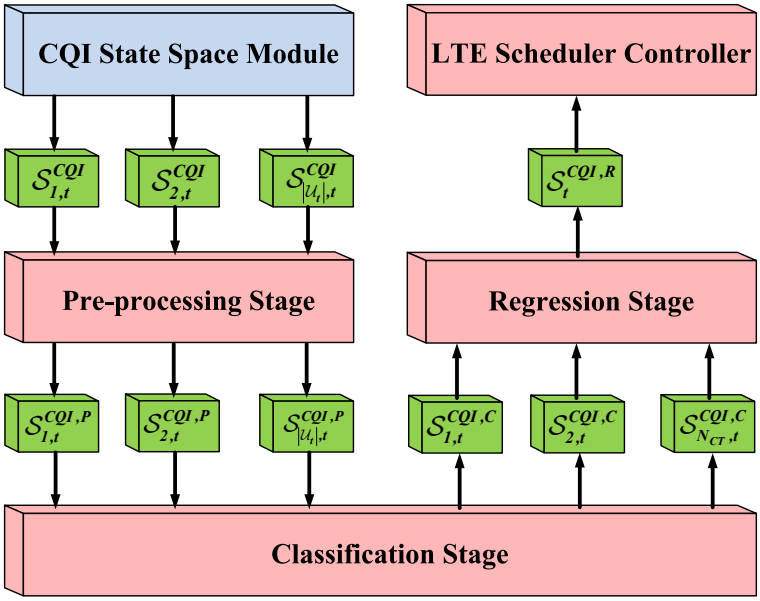


Fig. 4.3 The Proposed Architecture for the CQI State Space Aggregation

1. **The Preprocessing Stage**. The CQI state space $\mathcal{S}_t^{CQI}$ dimension is reduced from $D\left[\mathcal{S}_t^{CQI}\right]=|\mathcal{U}_t|\times|\mathcal{B}|$ to $|\mathcal{U}_t|\times N_{CQI}$, where N_{CQI} is the number of CQI instantaneous values (15 in LTE, 4 bits for uplink transmission). Basically, the full CQI report is replaced by a statistical report which is entitled ***CQI Mass Mode Report*** **(CQI-MMR)**. Mathematically, CQI-MMR can be represented as suggested in Equation 4.8:

$$\mathcal{S}_{i,t}^{CQI}=\bigcup_{j=1}^{|\mathcal{B}|}\left\{CQI_{i,j}[t]\right\}\xrightarrow{CQI-PS}\mathcal{S}_{i,t}^{CQI,P}=\bigcup_{v=1}^{N_{CQI}}\left\{MCQI_{i,v}[t]\right\},\ \forall i\in\mathcal{U}_t$$

$$\mathcal{S}_t^{CQI}=\bigcup_{i=1}^{|\mathcal{U}_t|}\mathcal{S}_{i,t}^{CQI}\xrightarrow{CQI-PS}\mathcal{S}_t^{CQI,P}=\bigcup_{i=1}^{|\mathcal{U}_t|}\mathcal{S}_{i,t}^{CQI,P}\qquad(4.8)$$

where, $\mathcal{S}_t^{CQI,P}$ is the preprocessed CQI state space and $MCQI_{i,v}[t]$ represents the mass mode CQI report. It is assumed that the CQI-PS stage is optional for some bandwidths (e.g. 1.4 to 3 MHz) as the original CQI state space has reasonable dimensions. In this case, the CQI report is entitled the ***CQI Normal Mode Report*** **(CQI-NMR).** For this study, only the CQI-MMR is considered for all simulation results.

2. **The Classification Stage**: The stage aims to reduce the preprocessed CQI state space $\mathcal{S}_t^{CQI,P}$ in order to avoid the redundant information that can be

reported by users with the same or appropriate CQI reports $\mathcal{S}_{i,t}^{CQI}$. One of the conditions is to have a very large collection of preprocessed CQI reports $\mathcal{S}_{i,t}^{CQI,P}$. Based on this collection, the set of the most representative CQI reports is chosen to represent the entire collection. The most representative reports are entitled *collection centers* denoted by the preprocessed CQI centers $\mathcal{S}_{i,t}^{CQI,CT}$. The input data is then classified based on these centers as belonging to one of the clusters. The impact of the CQI-CS in the CQI-PS is described by Eq. 4.9:

$$\mathcal{S}_{i,t}^{CQI,P} = \bigcup_{v=1}^{N_{CQI}} \left\{ MCQI_{i,v}[t] \right\} \xrightarrow{CQI-CS} \mathcal{S}_{k,t}^{CQI,CT} = \bigcup_{v=1}^{N_{CQI}} \left\{ MCQI_{k,v}[t] \right\}, \forall i \in \mathcal{U}_t$$

$$\mathcal{S}_{t}^{CQI,P} = \bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{S}_{i,t}^{CQI,P} \xrightarrow{CQI-CS} \mathcal{S}_{t}^{CQI,CT} = \bigcup_{k=1}^{N_{CT}} \mathcal{S}_{k,t}^{CQI,CT} \tag{4.9}$$

where N_{CT} represents the number of preprocessed CQI data centers. During the CQI-CS stage, the preprocessed CQI state space $\mathcal{S}_{t}^{CQI,P}$ dimension is reduced from $D\left[\mathcal{S}_{t}^{CQI,P}\right] = |\mathcal{U}_t| \times N_{CQI}$ to $D\left[\mathcal{S}_{t}^{CQI,CT}\right] = N_{CT} \times N_{CQI}$. By classifying the new entries, the space size $\left|\mathcal{S}_{t}^{CQI,CT}\right|$ of preprocessed CQI centers depends on the number of clusters N_{CT}. As observed, the preprocessed CQI center $\mathcal{S}_{k,t}^{CQI,CT}$ returns the index *k* for a considered input of preprocessed CQI observations. For the general purpose, the classified preprocessed CQI state space $\mathcal{S}_{t}^{CQI,C}$ is required in order to represent the percentage of CQI reports for different users located in different classes, where the dimension is $D\left[\mathcal{S}_{t}^{CQI,C}\right] = N_{CT}$. When the number of classes is very large (e.g. larger than 8), the regression stage is performed for the reduction of the classified preprocessed CQI state space $\mathcal{S}_{t}^{CQI,C}$.

3. **<u>The Regression Stage</u>**: The stage has the role of converting the classified preprocessed CQI state space $\mathcal{S}_{t}^{CQI,C}$ in statistical values. The result of the regression stage is directly provided as an input for the controller state space. Equation 4.10 follows the basic idea described above, where N_{pR} represents practically the dimension of the regressed state space.

Table 4.1 Methodologies for the CQI State Space Aggregation

Stage / Aggregation	1st Stage	2nd Stage	3rd Stage	Advantages	Disadvantages
Mass Mode	**R**	**No**	**No**	Eliminates the bandwidth length dependency	Statistical values are averaged over the number of RBs
Normal Mode	**N-R**	**No**	**No**	The distribution of RBs in a given bandwidth is known	Increases the CQI state space dimension
Classification	**No**	**R**	**No**	CQI state space does not depend on the system bandwidth	The number of classes has to be decided
Regression	**No**	**N-R**	**No**	Possible better precision	The dependency on the number of users is not avoided
Classification	**No**	**No**	**N-R**	Reduces the state space size for the regressed CQI states	Difficult to decide the number of classes
Regression	**No**	**No**	**R**	Better precision Reduces the CQI state dimension	The performance depends on the number of classes

$$\mathcal{S}_t^{CQI,C} = \bigcup_{k=1}^{N_{CT}} \mathcal{S}_{k,t}^{CQI,C} \xrightarrow{CQI-RS} \mathcal{S}_t^{CQI,R} = \bigcup_{pR=1}^{N_{pR}} \mathcal{S}_{pR,t}^{CQI,R} \tag{4.10}$$

To conclude, the CQI-PS is used to avoid the bandwidth dependency for the CQI state space $\mathcal{S}_t^{CQI}$, the CQI-CS avoids the number of users dependency for the preprocessed CQI state $\mathcal{S}_t^{CQI,P}$, and finally, the CQI-RS reduces the classified preprocessed CQI state space $\mathcal{S}_t^{CQI,C}$ to more comprehensive input values for the scheduler controller. Table 4.2 shows the possible advantages and disadvantages of the implied methodologies in the CQI state space aggregation. The dark green color represents the recommended approach which will be used in this research.

The proposed aggregation architecture follows the sequential structure in the sense that the output of each processing node is provided as an input to the next node. When excluding the preprocessing and regression nodes, the combination of supervised and unsupervised learning topologies is performed in the classification stage. More details about the proposed algorithms will be given in the following sub-sections. These nodes represent a descent order dependency where each node strongly depends on the previous one. Different operating modes must be decided in order to train the overall structure in a proper manner.

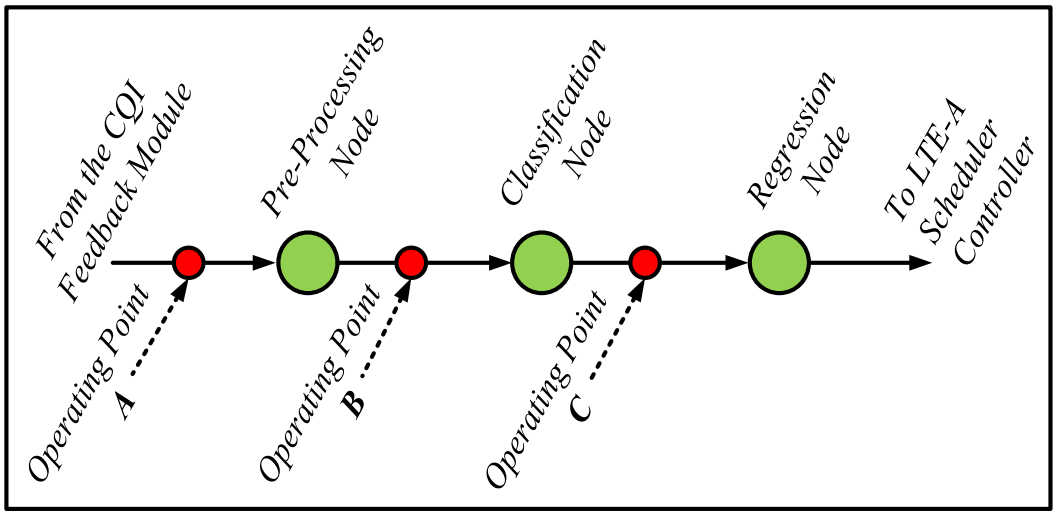


Fig. 4.4 Operating Points Involved in the CQI Aggregation Process

Figure 4.4 represents the dependencies among different nodes of the aggregation process. The operating points are depicted in red color. Each operating node is associated with an operating mode as shown in the Table 4.2. The exploration mode involves the training process of the classification stage. The validation mode represents a set of collected data which is used to validate the results from the training mode. The testing mode outputs the results to the next node. Therefore, the possible operating modes are presented below:

- ***Mode P:*0-*C:*0-*R:*0** – denotes the original implementation in which no aggregation function is applied at the output of the CQI Feedback module. This approach can be used in the case when several users are active in the cell (maximum 10 – typical femto-cell scenario).
- ***Mode P:*1-*C:*0-*R:*0** – the mass mode preprocessing stage is applied. It eliminates the dependency of the system bandwidth dimension.
- ***Mode P:*1-*C:*1-*R:*0** – when the preprocessing node is activated, the next procedure is to collect as many as possible preprocessed CQI data points.
- ***Mode P:*1-*C:*2-*R:*0** – based on the collected data points, the clustering algorithms can be applied in order to obtain the preprocessed CQI centers.
- ***Mode P:*1-*C:*3-*R:*0** – involves the training or the exploration process of the RBFNN structure considering the centers obtained in the working mode *P:*1-*C:*2- *R:*0 and the validation set from mode *P:*1-*C:*1-*R:*0.

Table 4.2. Operating Points Involved in the CQI Aggregation Process

Stage / Mode	Preprocessing Stage (P)	Classification Stage (C)	Regression Stage (R)
CQI Normal Mode	**P :0**		
CQI Mass Mode	**P :1**	**C :0**	
Collection CQI-MMR		**C :1**	
Clustering		**C :2**	
Exploration/Validation		**C :3**	
Exploitation		**C :4**	
Statistical Modeling			**R :1**

- ***Mode P:*1-*C:*4-*R:*0** – corresponds to the testing stage of the RBFNN classification procedure.
- ***Mode P:*1-*C:*4-*R:*1** – takes into consideration the CQI mass mode, the classification exploitation and the regression procedure. At this stage, the whole architecture is trained, validated and prepared to be exploited by other scheduler entities.

Before going through details about the CQI classification stage, Appendix B analyzes in details the cycle of the CQI report generation, considering both the link and system levels approaches. The impact of the propagation loss model is highlighted by using realistic scenarios accompanied with edifying results. It is very important to analyze the radio condition environment in order to apply the RBFNN classification under the most severe circumstances of the radio channels such as Jakes fast fading model.

In Appendix C is analyzed the CQI-PS stage in which the Top Mass CQI or Majority Mass CQI with reassignment principles are proposed in order to select the best CQI values in the computation of the preprocessed CQI state space $\mathcal{S}_t^{CQI,P,TM} \in \mathbb{R}_+^{N_{CQI}}$. Also, in Appendix C is provided a special algorithm which permits to collect the top/majority mass preprocessed CQI observations $\mathcal{S}_t^{CQI,P,TM}$ in order to determine the preprocessed CQI data centers for different bandwidths. The collection algorithm is stopped when the data set stays constant for a given time threshold. The clustering algorithms are performed based on the obtained set size $\left|\mathcal{S}_t^{CQI,P,TM}\right|$ for the collected top/majority mass preprocessed CQI reports.

4.5 Classification Stage in CQI Aggregation

In order to reduce the size of top/majority mass preprocessed CQI state space $\mathcal{S}_t^{CQI,P,TM}$ and to avoid the dependency of active bearers, the classification procedure is required to be performed in this sense. The general purpose is to find a general classification function $\mathcal{F}_C(\cdot)$ which maps at each TTI t the input preprocessed CQI state by using the top/majority mass modes from Appendix C in the classified preprocessed CQI state space $\mathcal{S}_t^{CQI,C}$ as shown in Eq. 4.11:

$$\mathcal{F}_C : \mathbb{R}^{|\mathcal{U}_t| \times N_{CQI}} \rightarrow \mathbb{R}^{N_{CT}} \qquad \mathcal{F}_C\left(\bigcup_{i=1}^{|\mathcal{U}_t|} \mathcal{S}_{i,t}^{CQI,P,TM}\right) = \mathcal{S}_t^{CQI,C} \tag{4.11}$$

where $N_{CT} = \left|\mathcal{S}_t^{CQI,CT}\right|$ is the number of k-means centers (classes) or the size of preprocessed CQI data centers. As mentioned earlier, the classified preprocessed CQI state $\mathcal{S}_t^{CQI,C}$ is a normalized space containing the percentages of preprocessed CQI states located in different k-means clusters. Let us define the size of the obtained state as $\left|\mathcal{S}_t^{CQI,C}\right| = \left(\left|\mathcal{U}_t\right|\right)^{N_{CT}}$. The number of classes should be chosen in order to balance the tradeoff between the classification accuracy and the system complexity. Therefore, the main difficulties which arise here refer to the determination of the number of centers N_{CT} and to the modalities of finding the optimal mapping function as defined in Eq. 4.11.

Finding the abovementioned mapping functions is a difficult task since such functions should approximate a given set of inputs into desire outputs without any predefined model. In other words, the classification function should be learned based on multiple input observations. In this sense, three approaches are extracted from the published literature, i.e., MLPNN [165], RBFNN [166-167] and Support Vector Networks or Support Vector Machine (SVM) [168].

One major drawback of the SVM is the fact that it was originally intended for the binary classification. Several methods aim to build multiclass classifiers based on different combinations of binary classifiers or using all possible classes at once [169], [170]. Another significant disadvantage of the SVM is the difficulty

of fine tuning a large set of parameters such as Kernel parameters involved in the SVM training processes [171]. In order to improve the prediction and the classification accuracy, the set of optimum parameters is determined based on different approaches such as simulated annealing, genetic algorithms or artificial fish swarm intelligence [171-173]. But the most important factor which makes the SVM unsuitable for CQI real-time classification refers to a much higher computational complexity for both exploration and exploitation stages when compared with other classifiers such as MLPNN or RBFNN [165], [174]. The computational cost of the SVM classification becomes even higher when the number of support vectors increase [175].

The most important features in the artificial neural networks (ANN) are the *learning* and the *generalization* [176]. The learning procedure approximates the output solution based on the training data set, whereas the generalization concept refers to the capability of accurate predictions based on validation data sets which are considered to be different from the training set. Two major problems are met under these circumstances such as:

- **The over-fitting symptom**: The trained network fits very well for the well-known training data set but offers a very poor generalization when the prediction (exploitation) step is used for the *unseen* observations. Similarly, the under-fitting effect highlights the insufficient learning based on the provided set of training observations [177].
- **Local minimum**: In the learning procedure, the objective is to find out the global minimum in the error minimization procedure between the expected and real outputs of the neural networks. Due to the error irregularities, the neural networks get stuck in local minima problem [177]. In other words, the local minimum is considered to be the global one.

Based on some studies, the SVM approaches overcome the ANN drawbacks [178], [179]. Clearly, for accurate classification and regression, RBFNN and MLPNN should use enhanced methods in order to avoid these issues. More details about the over-fitting and local minima avoidance in the CQI state space classification and regression are provided in the upcoming sub-sections.

RBFNN and MLPNN are both considered universal approximators and non-linear layered feed-forward neural networks [165]. From these reasons it is hard to predict which of these two approaches performs better for the preprocessed CQI state classification purpose. The MLPNN is viewed as a *stochastic approximation* whereas the RBFNN *aims to fit the curve for a high dimension input state space*. For the CQI classification problem, the RBFNN training aims to find the best fitting 15-dimensional surface. But the classification curve is obtained based on the well-known patterns. In Sub-section 4.5.2 is proposed an innovative concept in obtaining the preprocessed CQI state patterns which are necessary for a correct RBFNN classification.

The generalization aims to interpolate and to classify the testing preprocessed CQI vectors based on the trained surface. From the architectural point of view, the RBFNN is designed to have only three layers such as input, hidden and output, whereas the MLPNN can use more than one hidden layer. But the functionality inside of the hidden layers differs between the two. In the RBFNN case, the nodes in the hidden layers consist of *radial-basis functions* [180] which perform the non-linear transformation from the input space to the hidden space. Let us define the interpolation function $\mathcal{F}_o^{\mathcal{C}} : \mathbb{R}^{N_{CQI}} \to \mathbb{R}$, $\mathcal{F}_o^{\mathcal{C}}\left(\mathcal{S}_{i,t}^{CQI,P,TM}\right) = \mathcal{S}_{o,i,t}^{CQI,C}$ between the input nodes containing the preprocessed CQI reports for each user $i \in \mathcal{U}_t$ and the RBFNN output node $o = 1,..,N_O^{RBF}$ where N_O^{RBF} is the number of RBFNN output nodes. The RBFNN output state space for UE $i \in \mathcal{U}_t$ can be modeled such that $\mathcal{S}_{O,i,t}^{CQI,C} = \bigcup_{o=1}^{N_O^{RBF}} \mathcal{S}_{o,i,t}^{CQI,C}$. By discretizing the RBFNN output state for user $i \in \mathcal{U}_t$ and performing the binary transformation, the resulted number coincides with the center index of the preprocessed CQI state. More details about this novel idea are provided in Sub-section 4.5.2.

Based on the radial basis function, $\varphi_h : \mathbb{R}_+^{N_{CQI}} \to \mathbb{R}$ the same RBFNN interpolation function is calculated according to Eq. 4.12 [165]:

$$\mathcal{F}_o^{\mathcal{C}}\left(\mathcal{S}_{i,t}^{CQI,P,TM}\right) = \sum_{h=1}^{N_H^{RBF}} w_{h,o} \cdot \varphi_h\left(\left\|\mathcal{S}_{i,h,t}^{CQI,P,TM} - \mathcal{S}_{h,k}^{CQI,CT}\right\|\right) \tag{4.12}$$

where N_H^{RBF} is the number of RBFNN hidden nodes from the hidden layer, $w_{h,o}$ is the weight of hidden node *h* to the output node *o*, $\mathcal{S}_{h,k}^{CQI,CT}$ is the preprocessed CQI center $\forall k = 1,..,N_{CT}$ for the hidden node *h*, $\mathcal{S}_{i,h,t}^{CQI,P,TM}$ is the top/majority mass preprocessed CQI state at hidden layer *h* and finally, $\|\cdot\|$ is the Euclidian distance between the preprocessed CQI data points and the RBFNN centers. The RBFNN input state $\mathcal{S}_{i,h,t}^{CQI,P,TM}$ is the original input vector weighted by the first set of weights between the input and the hidden layer. Precise details about the RBFNN architecture are presented in Sub-section 4.5.2.

For a perfect interpolation, the number of RBFNN centers is considered to be equal to the number of hidden nodes $\left(N_{CT} = N_H^{RBF}\right)$. Then, the preprocessed CQI center at hidden node *h* $\mathcal{S}_{h,k}^{CQI,CT}$ becomes $\mathcal{S}_{k}^{CQI,CT}$. In order to decode the discretized version of the classified preprocessed CQI state for the RBFNN output layer $\mathcal{S}_{O,i,t}^{CQI,C}$ in a given center index, the number of output nodes becomes $N_O^{RBF} = log_2\left(N_H^{RBF}\right)$. Equation 4.12 is written under the matrix form as follows:

$$\begin{bmatrix} \mathcal{S}_{i,1}^{CQI,C} \\ \mathcal{S}_{i,ho}^{CQI,C} \\ \mathcal{S}_{i,N_O^{RBF}}^{CQI,C} \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,k} & w_{1,N_{CT}} \\ w_{ho,1} & w_{ho,k} & w_{ho,N_{CT}} \\ w_{N_O^{RBF},1} & w_{N_O^{RBF},k} & w_{N_O^{RBF},N_{CT}} \end{bmatrix} \cdot \begin{bmatrix} \varphi_1 \\ \varphi_k \\ \varphi_{N_{CT}} \end{bmatrix} \tag{4.13}$$

where $\mathcal{S}_{i,ho}^{CQI,C}$ represents the classified preprocessed CQI element at the input of the RBFNN output node where $ho = 1,..,N_O^{RBF}$.

Let us define the vector containing the radial basis functions for different hidden nodes $\Phi = \left\{\varphi_k,\ k = 1,..,N_{CT}\right\}^T$, the matrix of RBFNN hidden weights $W_O^{RBF} = \left\{w_{ho,k}\ , ho = 1,..,N_O^{RBF}, k = 1,..,N_{CT}\right\}$ and the classified preprocessed CQIs at the input of the RBFNN output layer $\mathcal{S}_{i.IO}^{CQI,C} = \left\{\mathcal{S}_{i,ho}^{CQI,C}, ho = 1,..,N_O^{RBF}, \forall i \in \mathcal{U}_t\right\}^T$. Equation 4.13 can be re-written under the following form:

$$\mathcal{S}_{IO}^{CQI,C} = W_O^{RBF} \cdot \Phi\left(\mathcal{S}_{i,t}^{CQI,P,TM}, \mathcal{S}^{CQI,CT}\right) \tag{4.14}$$

In order to obtain the classified preprocessed CQI state at the output RBFNN layer $\mathcal{S}_{O}^{CQI,C}$, the output activation function is applied to the classified preprocessed CQI state at the inputs of the RBFNN output layer $\mathcal{S}_{IO}^{CQI,C}$. More details about the RBFNN architecture are provided in Sub-section 4.5.2. Based on Micchelli's theorem, the vector of radial functions $\Phi(\cdot)$ is considered to be non-singular accepting its inverse matrix $\Phi^{-1}(\cdot)$[165]. Therefore, Equation 4.14 can be rewritten as shown in Equation 4.15:

$$W_{O}^{RBF} = \Phi^{-1}\left(\mathcal{S}_{i,t}^{CQI,P,TM}, \mathcal{S}^{CQI,CT}\right) \cdot \mathcal{S}_{IO}^{CQI,C} \tag{4.15}$$

In Equation 4.15, there are three *unknown* parameters, i.e., the classified preprocessed CQI state space at the inputs of the RBFNN output layer $\left(\mathcal{S}_{IO}^{CQI,C}\right)$, the set of preprocessed CQI centers $\left(\mathcal{S}^{CQI,CT}\right)$ and the RBFNN weights $\left(W_{O}^{RBF}\right)$. However, the RBFNN output space for each given set of inputs can be determined by using the feed-forward propagation technique which mathematically corresponds to Eq. 4.14. The set of RBFNN weights are corrected based on the back-propagation procedure which corresponds to Eq. 4.15. The set of parameters which needs to be determined in advance refers to the preprocessed CQI center space $\mathcal{S}^{CQI,CT}$. In this research, two different stages are proposed in training the RBFNN structure for the preprocessed CQI state space classification:

- **The First Stage** determines the set of preprocessed CQI centers based on the collected top/majority mass preprocessed CQI observations $\mathcal{S}_{i,t}^{CQI,P,TM}$, principle shown by the Algorithm C.2 in Appendix C. This stage is considered to be the ***unsupervised learning*** procedure, and the k-means clustering algorithms are used to obtain near-optimal sets of centers.
- **The Second Stage** trains (corrects) the RBFNN weights based on the feed-forward and backward training principles. This stage is considered to be the case of ***supervised learning***.

To conclude, the RBFNN training procedure in the CQI classification can be considered as a type of ***semi-supervised learning technique***.

4.5.1 Unsupervised Learning in CQI Classification

Given the collected 15-dimensional preprocessed CQI input space $\mathcal{S}_{i,t}^{CQI,P,TM} \in \mathbb{R}_{[0,1]}^{N_{CQI}}$, the problem is to determine a finite and optimal or near-optimal set of centers $\mathcal{S}^{CQI,CT} = \left\{ \mathcal{S}_k^{CQI,CT}, k = 1,..,N_{CT} \right\}$. If the size of the collected top mass preprocessed CQI state space $\left| \mathcal{S}_{i,t}^{CQI,P,TM} \right|$ is considered to be fixed to some large numbers depending on the system bandwidth, the abovementioned problem represents the typical case of geometric clustering algorithm [181]. The clustering algorithms aim to solve location problems and to find given sets of centers achieving different particular objectives, such as sum minimization of Euclidian distances from each data point to its nearest center point (k-median) [182] or the minimization of the maximum distance (k-center) [183]. Among these objectives, the minimization of the mean squared Euclidian distance from each data point to its neighbor center represents the most popular problem known as k-means clustering algorithm [184]. For the particular purpose of the CQI classification, the objective of k-means clustering is to find those centers in which the mean squared Euclidian distance from each point from the collected space of preprocessed top mass CQI states $\mathcal{S}_{i,t}^{CQI,P,TM}$ to the nearest center $\mathcal{S}_k^{CQI,CT}$ is minimized. The measure is known as ***squared-error distortion*** [185].

The squared-error distortion minimization represents the case of stochastic optimization problem in which the optimal solution is very difficult to be determined due to the clustering optimization problem characteristics. Therefore, the perfect location and the exact number of centers for the preprocessed CQI classification are very hard to be obtained in practice. Similar to the DSR-SMOO/CMOO problems, it is desirable to use some approximations based on heuristics algorithms which unfortunately cannot provide any kind of guarantee of the result quality. In this sense, some quality bounds should be decided for these heuristics [185]. Let us define a constant $c \geq 1$, an approximation value of the heuristic solution reported to the optimal one. Then, it is considered that the heuristic method can provide an approximation solution of $c \pm \varepsilon$, where ε represents the approximation error. For instance, in [182], the dynamic

programming based approach with adaptive hierarchical decomposition for the k-median problem achieves an approximation factor of $1+\varepsilon$ at the expense of the increased time complexity. It is important to notice that the approximation is achieved relatively to the local optimal solution rather than to the global one.

In the k-means clustering, the stochastic optimization problems can obtain the global optimum which depends on different heuristics algorithms. The heuristic algorithm calculates the distortion measure in time steps called *stages*. One of the most popular heuristics is Lloyd's algorithm [186], [187] which defines for each center point $\mathcal{S}_k^{CQI,CT}, k=1,...,N_{CT}$ the neighborhood data set in which the center is closest. The Lloyd's algorithm starts with the premise that if the obtained center lies near the neighborhood centroid, then the local minimum solution is guaranteed [188]. At each stage, the neighborhood of each center is computed and the center is moved to the centroid of the achieved neighborhood. This way, it is shown in [189] that Lloyd's algorithm converges to the local optimal solution. It is important to mention that the set of centers in Lloyd's case is not included in the collected preprocessed CQI set such as $\mathcal{S}^{CQI,CT} \not\subset \mathcal{S}_{i,t}^{CQI,P,TM}$.

Another solution for the k-means clustering is the Swap heuristic method [185], [190] in which the centers are swapped in and out from the set of candidate centers $\mathcal{S}_{Cand}^{CQI,CT}$ of the preprocessed CQI observations. At each stage, the swap heuristic improves the average distortion by removing one preprocessed CQI center $\mathcal{S}_r^{CQI,CT} \in \mathcal{S}^{CQI,CT}$, $\forall r=1,..,N_{CT}$ and replacing the remaining position with another center from the candidate list $\mathcal{S}_{a,Cand}^{CQI,CT} \in \mathcal{S}_{Cand}^{CQI,CT}$, $\forall a=1,..,N_k^{Cand}$, where the newest candidate CQI center becomes $\mathcal{S}_{a,Cand}^{CQI,CT} \in \mathcal{S}_{Cand}^{CQI,CT} \setminus \mathcal{S}_r^{CQI,CT}$. If the newest set of preprocessed CQI centers $\mathcal{S}_{New}^{CQI,CT} = \mathcal{S}^{CQI,CT} \bigcup \left\{\mathcal{S}_{a,Cand}^{CQI,CT}\right\} \setminus \left\{\mathcal{S}_r^{CQI,CT}\right\}$ performs better than the old one $\mathcal{S}_{Old}^{CQI,CT}$, then $\mathcal{S}_{New}^{CQI,CT}$ is saved and otherwise, the previous solution is restored. In [185] it is demonstrated that by performing a single swap heuristic at each stage will offer relatively poor approximation factor of about $\left(25+\varepsilon\right)$ with less computational complexity whereas the multiple swaps per each stage yields to an $\left(9+\varepsilon\right)$ approximation factor.

The Lloyd and Swap heuristics present certain complementary properties which are presented in the following:

- Lloyd can converge to the local optimal solution but has higher chances to get stuck in the local minima. Moreover, the most expensive cost is the computational complexity when the nearest neighbors for each center point are computed.
- Swap goes out from the local minima solution by swapping in and out new center points. On the other hand, it provides a larger approximation factor when compared to Lloyd.

The idea of Lloyd's algorithm is to perform under some iterations. For each iteration, the set of preprocessed CQI centers $\mathcal{S}^{CQI,CT}$ is randomly chosen and then, the neighborhood of each center is calculated. At each stage the centroid is computed and the center is moved to the centroid. The algorithm is repeated at each stage until the consecutive average distortion falls below a given threshold. The number of stages until the Lloyd algorithm meets the convergence criterion is called *run*. At the beginning of each run, a new set of centers is randomly chosen and the algorithm is repeated in the same manner. However, the set of centers which provides the minimum distortions is saved. Ideally would be if at each stage, the Lloyd and Swap heuristics are combined in order to escape from the local minima problem. Based on the above idea, in [185] the Simulated Annealing (SA) approach [191] is proposed to achieve a dynamic combination of Lloyd and Swap methods in order to determine the local optimum set of centers. However, even with the SA method, the local minima avoidance is not guaranteed. In order to solve this drawback, a new method such as the ***Simulated Annealing with Stochastic Tunneling (SAST)*** is proposed, and the concepts about the stochastic functions are provided in [192], [193], [194]. The details of the proposed heuristic model for the determination of the preprocessed CQI centers are given in Sub-section 4.5.1.4. One of the most problematic issues for these heuristic approaches refers to the very high computation cost when the neighborhoods of each center are calculated. This work aims to store data in the k-d tree [195] and the neighborhood calculation is based on the filtering algorithm proposed in [196].

The obtained centers under the dynamic SAST method are used by the RBFNN structure to generalize the uncontrollable CQI state space classification. One major problem is to determine the optimal number of *k* centers for each LTE bandwidth since the preprocessing stage is affected by different system bandwidth configurations. Then, the purpose of the k-means clustering usage is to provide six sets of optimal centers with the lowest distortion for each possible bandwidth [152] that can be used in the classification procedure. It is important to mention that even if the set of preprocessed CQI centers $\mathcal{S}^{CQI,CT}$ and the number of centers N_{CT} are determined as offline processes, the number of centers decides the number of RBFNN hidden nodes which affects the online CQI classification and the regression functions. The minimum number of centers which indicates a sufficiently small average distortion should be chosen.

4.5.1.1 The Filtering Principle for Calculating the Preprocessed CQI Centers

For the purpose of the implementation efficiency, the 15-dimensional preprocessed top/majority mass CQI state space is stored in the k-d tree [195]. The top/majority mass modes of the preprocessed CQI observations are achieved based on the reassignment methods (Appendix C). More precise details about the k-d tree can be found in [195]. The k-d tree is considered to be a binary tree representation. For the particular case of 15-dimension preprocessed CQI space, the k-d tree defines a *box* which is in fact a 15-dimensional hyper-rectangle. The *bounding box* is the largest box which contains a number of $\left|\mathcal{S}_{i,t}^{CQI,P,TM}\right|$ data points from the collected top/majority mass preprocessed CQI observations. Each point from the collection of the preprocessed top/majority CQI observations $\mathcal{S}_{i,t}^{CQI,P,TM}$ defines its own box called *cell*. The k-d binary tree splits the bounding box starting from the cell root in two closed boxes which are considered to be two-axis-orthogonal hyper-rectangle. The process continues in a similar way for each of the existing cells until each box contains at least one point. In this case, the point is called *leaf*. All data points contained by one cell in the descending sense of the k-d construction are considered to be the associated points.

Briefly described below is the filtering algorithm which was initially proposed in [196] and it is applied in this study for the calculation of the preprocessed CQI centers. Let us define the internal k-d tree for the top/majority preprocessed CQI data node $\mathcal{S}_{i,u,t}^{CQI,P,TM}$, where $u = 1,..,\left|\mathcal{S}_{i,t}^{CQI,P,TM}\right|$ and $\mathcal{V}_{i,u,t}^{CQI,P,TM}$ are the points in its associated cell. For each preprocessed CQI node $\mathcal{S}_{i,u,t}^{CQI,P,TM}$, the number of associated points $\left|\mathcal{V}_{i,u,t}^{CQI,P,TM}\right|$ and the centroid $\mathcal{C}_{\mathcal{V},u}^{CQI,P,TM}$ for all associated preprocessed CQI points in a given cell are calculated. Then, the weighted centroid is $\mathcal{WC}_{\mathcal{V},u}^{CQI,P,TM} = \mathcal{C}_{\mathcal{V},u}^{CQI,P,TM} / \left|\mathcal{V}_{i,u,t}^{CQI,P,TM}\right|$. This computation stage is very important since in the Lloyd stage, for a given preprocessed CQI center $\mathcal{S}_{k}^{CQI,CT}$, the centroid of the nearest data points is already calculated, and the new center becomes $\mathcal{S}_{k,New}^{CQI,CT} = \mathcal{WC}_{\mathcal{V},u}^{CQI,P,TM}$ for a given top/majority mass preprocessed CQI input $\mathcal{S}_{i,u,t}^{CQI,P,TM}$ of the k-d tree node.

As mentioned earlier, for each data point of the k-d tree, the set of center candidates $\mathcal{S}_{Cand,u}^{CQI,CT}$ is saved, assuming that some data points from the cell $\mathcal{V}_{i,u,t}^{CQI,P,TM}$ are closest to one of the candidate center rather than to the centroid $\mathcal{C}_{\mathcal{V},u}^{CQI,P,TM}$ which is considered the midpoint of cell $\mathcal{V}_{i,u,t}^{CQI,P,TM}$. The preprocessed CQI center $\mathcal{S}_{k}^{*CQI,CT} \in \mathcal{S}_{Cand,u}^{CQI,CT}$ which is closest to the centroid $\mathcal{C}_{\mathcal{V},u}^{CQI,P,TM}$ is selected and the rest of centers $\mathcal{S}_{Cand,u}^{CQI,CT} \setminus \left\{\mathcal{S}_{k}^{*CQI,CT}\right\}$ from the candidate list can be filtered only and only if there is no other point from the set of associated points $\mathcal{V}_{i,u,t}^{CQI,P,TM}$ which is closer to other preprocessed CQI center $\mathcal{S}_{kc}^{CQI,CT} \in \mathcal{S}_{Cand,u}^{CQI,CT}, \forall kc \neq k$ than to $\mathcal{S}_{k}^{*CQI,CT}$. More details about the center filtering procedure can be found in [196].

4.5.1.2 The Iterated Lloyd Algorithm

One of the most important problems in k-means clustering is the initial selection of the set of preprocessed CQI centers $\mathcal{S}^{CQI,CT}$ from the collection of the preprocessed top/majority mass CQI states $\mathcal{S}_{i,u}^{CQI,P,TM}$. The Lloyd algorithm does not specify the rule of the center selection procedure. Therefore, the random

initialization of the center set can be performed in this sense. The iterated Lloyd algorithm performs in time steps called stages. At each stage, the neighborhood set for each center is determined based on the filtering algorithm presented in the previous sub-section. Let us define the preprocessed CQI data set which is located in the neighborhood of the center $\mathcal{S}_{k,t_{st}}^{CQI,CT}$ such as $\mathcal{N}_{k,kc,t_{st}}^{CQI,P,TM}$ at stage t_{st} where $t_{st}=1,...,N_{St}^{max}$ represents the stage index from the total number of N_{st}^{max} stages. For each neighborhood set of each center at stage t_{st}, the weighted centroid is calculated according to Eq. 4.16 [200]:

$$\mathcal{WC}_{k,kc,t_{st}}^{CQI,P,TM}=1\Big/\left|\mathcal{N}_{k,kc,t_{st}}^{CQI,P,TM}\right|\cdot\sum\nolimits_{n=1}^{\left|\mathcal{N}_{k,kc}^{CQI}\right|}\mathcal{S}_{n,t_{st}}^{CQI,P,TM} \tag{4.16}$$

where $\mathcal{WC}_{k,kc,t_{st}}^{CQI,P,TM}$ is the weighted centroid for the top/majority mass preprocessed CQI observations represented by the center $\mathcal{S}_{k,t_{st}}^{CQI,CT}$ with the candidate center $\mathcal{S}_{kc,t_{st}}^{CQI,CT}\subset\mathcal{N}_{k,kc,t_{st}}^{CQI,P,TM}$ at stage t_{st}. At the end of each stage, the new preprocessed CQI center set becomes the weighted centroid of its neighborhood such that $\mathcal{S}_{k,kc,t_{st}}^{CQI,CT}\leftarrow\mathcal{WC}_{k,kc,t_{st}}^{CQI,P,TM}$. After this assignment, the algorithm performs the next stage $t_{st}+1$. It is decided to use the notation of $\mathcal{S}_{k,kc,t_{st},1}^{CQI,CT}=\mathcal{S}_{k,t_{st}}^{CQI,CT}$, denoting the center point *k* with the candidate index *kc* at stage t_{st} for the iterated Lloyd algorithm. The iterated Lloyd algorithm aims to minimize at each stage the total distortion from each data point $\mathcal{S}_{i,u,t_{st}}^{CQI,P,TM}$ to its nearest center $\mathcal{S}_{k,kc,t_{st},1}^{CQI,CT}$. Therefore, the expanded Lloyd stochastic optimization problem can be defined as follows:

$$\begin{aligned}
&(P_L):\min_{\pi_L[t_{st}]}\sum_{u=1}^{\left|\mathcal{S}_{i,u}^{CQI}\right|}\sum_{k=1}^{N_{CT}}w_{u,k}^{L,1}[t_{st}]\cdot\sum_{kc=1}^{N_{CT,k}^{Cand}}w_{k,kc}^{L,2}\cdot\left\|\mathcal{S}_{i,u,t_{st}}^{CQI,P,TM}-\frac{1}{\left|\mathcal{N}_{k,kc}^{CQI}\right|}\sum_{n=1}^{\left|\mathcal{N}_{k,kc}^{CQI}\right|}w_{kc,n}^{L,3}[t_{st}]\cdot\mathcal{S}_{n,t_{st}}^{CQI,P,TM}\right\|\\
&\sum_{k=1}^{N_{CT}}w_{u,k}^{L,1}[t_{st}]=1,\ u=1,...,\left|\mathcal{S}_{i,u}^{CQI,P,TM}\right|\\
&(C_L):s.t.\ \sum_{kc=1}^{N_{CT,k}^{Cand}}w_{k,kc}^{L,2}[t_{ph}]=1,\ k=1,...,N_{CT}\\
&\sum_{n=1}^{\left|\mathcal{N}_{k,kc}^{CQI}\right|}w_{kc,n}^{L,3}[t_{st}]=1,...,\left|\mathcal{N}_{k,kc}^{CQI}\right|,\ kc=1,...,N_{CT,k}^{Cand}\\
&w_{k,u}^{L,1}[t_{st}],w_{k,kc}^{L,2}[t_{ph}]\in\{0,1\},w_{kc,n}^{L,3}[t_{st}]\in\left\{1,...,\left|\mathcal{N}_{k,kc}^{CQI}\right|\right\}
\end{aligned} \tag{4.17}$$

where $w_{u,k}^{L,1}\left[t_{st}\right]$ assigns a center for each preprocessed CQI observation, $w_{k,kc}^{L,2}\left[t_{ph}\right]$ associates a candidate center *kc* to each current center *k* at each phase stage and finally, $w_{kc,n}^{L,3}\left[t_{st}\right]$ assigns associated points to each preprocessed CQI candidate center. The optimization function from $\left(P_L\right)$ is concave and the set of constraints $\left(C_L\right)$ is non-convex [189]. Moreover, the exposed problem is a typical case of non-linear integer programming problem. Therefore, the global optimum solution is not guaranteed. The major problem is represented by the fact that $\left(P_L\right)$ cannot be the subject of linearization due to the product $w_{u,k}^{L,1}\left[t_{st}\right]\cdot w_{k,kc}^{L,2}\left[t_{ph}\right]\cdot w_{kc,n}^{L,3}\left[t_{st}\right]$. The set of candidate centers is always chosen from the total preprocessed CQI data set $\mathcal{S}_{i,t}^{CQI,P,TM}$ which involves the non-convexity property. Then, the iterated Lloyd is considered to be a local optimum search algorithm which aims to find the best set of centers by using local iterations with random restarts of the center set.

Let us define $\mathcal{S}_{Best}^{CQI,CT}$ the best set of preprocessed CQI centers which is detected so far. If the set of preprocessed CQI centers $\mathcal{S}_{t_{st}}^{CQI,CT}$ at stage t_{st} achieves a better average distortion $\overline{D}\left[t_{st}\right]$ when compared against the distortion being obtained for the set of centers $\mathcal{S}_{Best}^{CQI,CT}$, then the best set of preprocessed CQI centers becomes $\mathcal{S}_{Best}^{CQI,CT}=\mathcal{S}_{t_{st}}^{CQI,CT}$, and the reasoning on this is illustrated by the following equation:

$$\mathcal{S}_{Best}^{CQI,CT}=\begin{cases}\mathcal{S}_{t_{st}}^{CQI,CT}, if\,\overline{D}_{t_{st}}<\overline{D}_{Best}\\ \mathcal{S}_{Best}^{CQI,CT}, if\,\overline{D}_{t_{st}}>\overline{D}_{Best}\end{cases} \tag{4.18}$$

where the average distortion at stage t_{st} is $\overline{D}\left[t_{st}\right]=1\Big/\left|\mathcal{S}_i^{CQI,P,TM}\right|\cdot\sum_{u=1}^{\left|\mathcal{S}_{i,u}^{CQI}\right|}D_u\left[t_{st}\right]$, and the instantaneous distortion for a given preprocessed CQI data point and its corresponding center is calculated by using Eq. 4.19 [200]:

$$D_u\left[t_{st}\right]=\left\|\mathcal{S}_{i,u,t_{st}}^{CQI,P,TM}-\mathcal{S}_{k,kc,t_{st},1}^{CQI,CT}\right\|=\sum_{d=1}^{N_{CQI}}\left(MCQI_{i,u,t_{st}}^{P,TM}\left[d\right]-MCQI_{k,kc,t_{st},1}^{CT}\left[d\right]\right)^2 \tag{4.19}$$

where $\|\cdot\|$ represents the Euclidian distance between two 15-dimensional points.

By using the random choice of centers only at once, the Lloyd algorithm gets stuck in the local minima solution, being impossible to find a better set of preprocessed CQI centers than $\mathcal{S}_{Best}^{CQI,CT}$. For this reason, it is preferable to use multiple random restarts for the sets of data centers. The problem which arises here refers to the moment in the sequence of stages in which the initial set of centers should be changed. In this sense, a quality measure is defined in terms of the Accumulated Relative Distortion Loss (ARDL) expressed by Eq. 4.20 [196]:

$$RDL_A\left[t_{st}\right] = \left(D_{INIT} - D\left[t_{st}\right]\right) \Big/ D_{INIT} \tag{4.20}$$

where $D\left[t_{st}\right] = \sum_{u=1}^{\left|\mathcal{S}_{i,u}^{CQI}\right|} D_u\left[t_{st}\right]$ is the overall distortion for each data point $\mathcal{S}_{i,u,t_{st}}^{CQI,P,TM}$ from the top/majority mass preprocessed CQI state space, and D_{INIT} represents the initial distortion. From Equation 4.20, it is easy to observe that when the accumulated relative distortion is $RDL_A\left[t_{st}\right] > 0$, the current center set provides a lower distortion when compared with the initial one. A number of stages is grouped in *drops* or *runs* where $t_{run}^{max} = N_{st/run}^{max} \cdot t_{st}$ and $N_{st/run}^{max}$ is an input parameter which represents the maximum number of stages included in one run epoch. A number of runs is grouped in phases. The beginning of a new phase implies that the overall set of centers is exchanged with other points from the candidate list by setting the assignment variable of the candidate centers for the Lloyd algorithm such as $w_{k,kc}^{L,2}\left[t_{ph}\right] = 1$, where $t_{ph} = N_{runs/ph}\left[t_{ph}\right] \cdot t_{ph}^{0}$ and $N_{runs/ph}\left[t_{ph}\right]$ represents the run counter in a given phase and t_{ph}^{0} is the initial time step when a new phase starts. A run epoch is successfully finished when the relative distortion becomes $RDL_A\left[t_{st}\right] > RDL_A^{Min}, \forall t_{run}^{Succ} = \left(1,...,N_{st/run}^{max}\right) \cdot t_{st}$ where t_{run}^{Succ} is the run epoch, moment when it is successfully finished. In this case, the set of preprocessed CQI centers is not randomized and the initial distortion becomes $D_{INIT} = D\left[t_{run}^{Succ}\right]$. In the case of unsuccessful termination, a new phase starts, a new center set is randomized from the collected set of CQI observations and then, the initial distortion becomes: $D_{INIT} = D\left[t_{ph}\right]$, where $\forall t_{ph} = N_{runs/ph}\left[t_{run}\right] \cdot t_{run}^{Unsucc}$. For a comprehensive explanation, Algorithm 4.1 highlights the aspects discussed above.

Algorithm 4.1 Iterated Lloyd Heuristic

Require: run^{New} : boolean value which denotes if there is a new run epoch

ph^{New} : boolean value which denotes if there is a new phase epoch

1. **while** $\left(t_{st} \le N_{st}^{max}\right)$
2. **start** stage t_{st}
3. **if** $\left(run^{New} = true\right)$
4. $run^{New} = false$, $t_{run}^{st} = 0$
5. **end if**
6. **if** $\left(ph^{New} = true\right)$
7. Randomize set of centers: $\mathcal{S}_{k,kc,t_{st},1}^{CQI,CT} \quad \forall k = 1,...,N_{CT}$, $\forall kc = 1,...,N_{CT,k}^{Cand}$
8. $ph^{New} = false$, $D_{INIT} = D\left[t_{ph}\right]$
9. **end if**
10. **for** each center $\mathcal{S}_{k,kc,t_{st},1}^{CQI,CT}$
11. **Get neighbors of** $\mathcal{S}_{k,kc,t_{st},1}^{CQI,CT}$: $\mathcal{N}_{i,kc,t_{st}}^{CQI,P,TM}$
12. **Filter Centers** $\left(\mathcal{S}_{k,kc,t_{st},1}^{CQI,CT}, \mathcal{N}_{i,kc,t_{st}}^{CQI,P,TM}\right)$
13. **Determine centroid** $\mathcal{WC}_{k,kc,t_{st}}^{CQI,CT}$ **based on Eq. (4.16)**
14. **Move centroid to center:** $\mathcal{S}_{k,kc,t_{st}}^{CQI,P,CT} \leftarrow \mathcal{WC}_{k,kc,t_{st}}^{CQI,P,TM}$
15. **end for**
16. $t_{run}^{st} = t_{run}^{st} + 1$
17. **if** $\left(t_{run}^{st} < N_{st/run}^{max}\right)$
18. **if** $\left(RDL_A\left[t_{run}\right] > RDL_A^{Min}\right)$
19. $run^{New} = true$, $ph^{New} = false$
20. $D_{INIT} = D\left[t_{run}\right]$
21. **end if**
22. **else if** $\left(t_{run}^{st} = N_{st/run}^{max}\right)$
23. $run^{New} = true$
24. **if** $\left(RDL_A\left[t_{run}\right] > RDL_A^{Min}\right)$
25. $ph^{New} = false$
26. $D_{INIT} = D\left[t_{run}\right]$
27. **else if**
28. $ph^{New} = true$
29. **end if**
30. **end if**
31. **if** $\left[\left(\overline{D}\left[t_{st}\right] < \overline{D}_{Best}\right) \&\& \left(run^{New} = true\right) \&\& \left(ph^{New} = false\right)\right]$
32. $\mathcal{S}_{Best}^{CQI,CT} = \mathcal{S}_{t_{st}}^{CQI,CT}$
33. **end if**
34. **end** stage t_{st}
35. **end while**

4.5.1.3 The Single Swap Heuristic Algorithm

The swap heuristic method does not consider the centroid calculation at each stage being able to improve the best set of centers $\mathcal{S}_{Best}^{CQI,CT}$ when compared with Lloyd due to its ability to remove one random center from the center space $\mathcal{S}_{k,kc,t_{st},2}^{CQI,CT} \in \mathcal{S}_{t_{st},2}^{CQI,CT}$ with one center from the candidate list of centers such as $\mathcal{S}_{kc,t_{st},2}^{CQI,CT} \in \mathcal{S}_{Cand}^{CQI,CT}, \forall kc \neq k$. An advantage of the swap heuristic is to avoid the blocking effect into the local minima but it may take a longer time to find an optimal set of centers. One of the biggest concerns in the swap heuristic is the decision of the set of preprocessed CQI candidate centers $\mathcal{S}_{Cand}^{CQI,CT}$. For the this study, it is decided to use the set of candidate centers being defined as follows: $\mathcal{S}_{kc,Cand}^{CQI,CT} = \mathcal{S}_{i,t_{st}}^{CQI,P,TM} \setminus \mathcal{S}_{k,kc,t_{st},2}^{CQI,CT}, \forall k \neq kc$. This approach may take a longer time of running until it finds the optimal centers but has the biggest advantage of escaping from the local minima problem. Then, the optimization problem of the swap heuristic algorithm becomes:

$$
\begin{aligned}
(P_S):\ & \min_{\pi_S[t_{st}]} \sum_{u=1}^{\left|\mathcal{S}_{i,u}^{CQI}\right|} \sum_{k=1}^{N_{CT}} w_{u,k}^{S,1}[t_{st}] \cdot \sum_{kc=1}^{N_{CT,k}^{Cand}} w_{k,kc}^{S,2}[t_{st}] \cdot \left\| \mathcal{S}_{i,u,t_{st}}^{CQI,P,TM} - \mathcal{S}_{k,kc,t_{st},2}^{CQI,P,TM} \right\| \\
& \sum_{k=1}^{N_{CT}} w_{u,k}^{S,1}[t_{st}] = 1,\ u = 1,\ldots,\left|\mathcal{S}_{i,u}^{CQI,P,TM}\right| \\
(C_S):\ & s.t.\ \sum_{kc=1}^{N_{CT,k}^{Cand}} w_{k,kc}^{S,2}[t_{ph}] = 1,\ k = 1,\ldots,\left|\mathcal{S}_{i,u}^{CQI}\right|, \forall k \neq kc \\
& w_{u,k}^{S,1}[t_{st}], w_{k,kc}^{S,2}[t_{st}] \in \{0,1\}
\end{aligned}
\tag{4.21}
$$

where $w_{k,kc}^{S,2}[t_{st}]$ is the assignment variable which associates a candidate center *kc* to each current center *k* and it can be changed at each stage for each center at once and $w_{u,k}^{S,1}[t_{st}]$ is the assignment variable which associates a center for each top/majority mass preprocessed CQI observation. For the current purpose, the maximum number of swaps N_{SW}^{max} is defined per each running epoch. The properties of Eq. 4.21 remain unchanged when compared to Eq. 4.17 due to the fact that the (P_S) minimization function is concave and the constraint set (C_S) is non-convex. Algorithm 4.2 briefly explains the single swap heuristic reasoning.

Algorithm 4.2 Single Swap Heuristic

Require: run^{New} : boolean value which denotes if there is a new run epoch
ph^{New} : boolean value which denotes if there is a new phase epoch

1. while $\left(t_{st} \le N_{st}^{max}\right)$
2. **start** stage t_{st}
3. **if** $\left(run^{New} = true\right)$ $run^{New} = false$, $t_{run}^{st} = 0$
4. **end if**
5. **if** $\left(ph^{New} = true\right)$
6. Randomize the entire set of centers: $\mathcal{S}_{t_{st},2}^{CQI,CT}$, $ph^{New} = false$
7. **else**
8. Swap center one *k* with candidate *kc* : $\mathcal{S}_{k,kc,t_{st},2}^{CQI,P,CT}$
9. $\forall k \neq kc, \forall k = \{1,...,N_{CT}\}, \forall kc = \{1,...,N_{CT,k}^{Cand}\}$
10. **end if**
11. $t_{run}^{st} = t_{run}^{st} + 1$
12. **if** $\left(t_{run}^{st} < N_{st/run}^{max}\right)$
13. **if** $\left(N_{SW} = N_{SW}^{max}\right)$ $ph^{New} = true$, $run^{New} = true$
14. **end if**
15. **else if** $\left(t_{run}^{st} = N_{st/run}^{max}\right)$ $run^{New} = true$, $ph^{New} = true$
16. **end if**
17. **if** $\left(\overline{D}[t_{st}] < \overline{D}_{Best}\right)$ $\mathcal{S}_{Best}^{CQI,P,CT} = \mathcal{S}_{t_{st}}^{CQI,P,CT}$
18. **else** $\mathcal{S}_{t_{st}}^{CQI,P,CT} = \mathcal{S}_{Best}^{CQI,P,CT}$
19. **end if**
20. **end** stage t_{st}
21. **end while**

4.5.1.4 Lloyd-Swap Heuristics Based Simulated Annealing with Stochastic Tunneling

As mentioned above, the static Lloyd minimizes the average distortion between the collected preprocessed CQI data points and the corresponding set of obtained centroids under the local minima problem. For instance, when a run epoch finishes in success, the initial distortion D_{INIT} takes the value of the previous distortion value and the successful set of centers is saved as the best solution. The question to be asked here is "***are these centers able to drive the searching space to the global minimum solution*?**" The global minimum solution can be reached if the centers from the candidate list and from the current set of

centers are swapped in and out at different stages. The question which arises here is "***how many stages the swap heuristic needs in order to guarantee the optimal minimum solution*?**" It is difficult to find the answers for both questions since both optimization problems are considered NP hard. Therefore, the architecture obtained by combining both Lloyd and Swap approaches can take the advantages of both approaches and it can minimize at the same time, the disadvantages. The main advantage considered here refers to the probability of finding a better local minima solution when compared with the static versions of both heuristics. The obtained stochastic optimization problem is presented by Eq. 4.22:

$$
\begin{aligned}
(P_{SAST}):\ & \underset{\pi_{SAST}[t_{st}]}{min} \sum_{h=1}^{2}\sum_{u=1}^{\left|\mathcal{S}_{i,u}^{CQI}\right|} w_{l,u}^{SAST,1}[t_{st}]\cdot\sum_{k=1}^{N_{CT}} w_{u,k}^{SAST,2}[t_{st}]\cdot\sum_{kc=1}^{N_{CT,u}^{Cand}} w_{k,kc}^{SAST,3}[t_{st}]\cdot\left\|\mathcal{S}_{i,u}^{CQI,P,TM}-\mathcal{S}_{kc,l}^{CQI,CT}\right\| \\
& \sum_{l=1}^{2} w_{l,u}^{SAST,1}[t_{st}]=1,\ u=1,\ldots,\left|\mathcal{S}_{i,u,t_{st}}^{CQI,P,TM}\right| \\
(C_{SAST}):\ s.t.\ & \sum_{k=1}^{N_{CT}} w_{u,k}^{SAST,2}[t_{st}]=1,\ u=1,\ldots,\left|\mathcal{S}_{i,u}^{CQI,P,TM}\right| \\
& \sum_{kc=1}^{N_{CT,k}^{Cand}} w_{k,kc}^{SAST,3}[t_{st}]=1,\ \forall k=1,\ldots,\left|\mathcal{S}_{t_{st}}^{CQI,CT}\right| \\
& w_{l,u}^{SAST,1}[t_{st}], w_{u,k}^{SAST,2}[t_{st}], w_{k,kc}^{SAST,3}[t_{st}]\in\{0,1\}
\end{aligned}
\tag{4.22}
$$

where $w_{l,u}^{SAST,1}[t_{st}]$ is the assignment variable which selects Lloyd or the single swap heuristic methodologies at each stage, the variable $w_{u,k}^{SAST,2}[t_{st}]$ assigns a center point to each preprocessed CQI observation and the variable $w_{k,kc}^{SAST,3}[t_{st}]$ selects a candidate center for each center *k*. The objective of the proposed method is to obtain a good approximation of the global minimum solution such that $(O_{SAST}):\ \overline{D}^{*}[t_{ts}]<\overline{D}_{Best}$, where $\overline{D}_{Best}$ represents the best average distortion discovered so far and $\overline{D}^{*}$ represents a better distortion value being reached at stage t_{st} which approximates the global optimal solution. The optimization problem P_{SAST} is nonlinear due to the product between assignment variables at each stage t_{st} $w_{l,u}^{SAST,1}[t_{st}]\cdot w_{u,k}^{SAST,2}[t_{st}]\cdot w_{k,kc}^{SAST,3}[t_{st}]$. To conclude, three decisions can be taken based on the optimization problem from Eq. 4.22:

1. Accept or refuse the current non-better solution such as the average distortion at stage t_{st} $\overline{D}[t_{ts}]$. If the current solution is accepted and the best average distortion becomes $\overline{D}_{Best} = \overline{D}[t_{ts}]$, then the obtained set of centers is saved.
2. Lloyd or Swap decision $\left(w_{l,u}^{SAST,1}[t_{st}]\right)$ in the current stage;
3. Decide based on $\left(w_{u,k}^{SAST,2}[t_{st}], w_{k,kc}^{SAST,3}[t_{st}]\right)$ to perform another Lloyd search in the current stage rather than going to the newest solution acceptance.

All these decisions are adopted in order to reach the global minimum solution for the preprocessed CQI centers $\mathcal{S}_{t_{st}}^{*CQI,CT}$ with the optimal average distortion $\overline{D}^{*}[t_{ts}]$. The simulated annealing is known as a powerful meta-heuristic methodology in finding a global optimal solution approximation for stochastic optimization problems when the searching space is very large [191]. Recent studies indicate that other meta-heuristics are able to perform better than SA in many domains [192], [193], [194].

The ***Stochastic Tunneling*** (ST) approach is one of such methodologies which improve classical SA approach by using Monte Carlo samplings in the objective minimization [192]. Let us define the average distortion for a given set of centers $\mathcal{SC}_{tk,t_{st}}^{CQI,CT} = \mathcal{S}_{t_{st}}^{CQI,CT}$ at stage t_{st} $\overline{D}\left(\mathcal{SC}_{tk,t_{st}}^{CQI,CT}\right)$, where $tk = 1,...,\left|\mathcal{SC}^{CQI,CT}\right|$ represents the center set index from a total number of center sets $\left|\mathcal{SC}^{CQI,CT}\right|$ with N_{CT} number of centers for each set. The SAST aims to solve the stochastic optimization problem $\left(P_{SAST}\right)$, starting from any initial state $\mathcal{SC}_{0}^{CQI,CT}$ to the optimal set of CQI centers $\mathcal{SC}_{tk}^{*CQI,CT}$ with the minimum average distortion $\overline{D}^{*}$.

For each state of center set $\mathcal{SC}_{tk}^{CQI,CT}$, a neighborhood function such as $\mathcal{N}\left(\mathcal{SC}_{tk}^{CQI,CT}\right)$ is given, where the center set index is $tk = 1,...,\left|\mathcal{SC}^{CQI,CT}\right|$. For the optimization problem $\left(P_{SAST}\right)$ purpose, SAST uses two probabilistic decisions of accepting Swap or Lloyd for the next stage $\left(w_{l,u}^{SAST,1}[t_{st}]\right)$ and accepting or refusing

non-better neighborhood solution with the average distortion $\overline{D}'$ in the center state $\mathcal{SC}_{tk}^{'CQI,CT} \in \mathcal{N}\left(\mathcal{SC}_{tk,Best}^{CQI,CT}\right)$ when the best distortion solution which is found so far is $\overline{D}_{Best}\left(\mathcal{SC}_{tk,Best}^{CQI,CT}\right)$ by controlling the decision variables $\left(w_{u,k}^{SAST,2}\left[t_{st}\right], w_{k,kc}^{SAST,3}\left[t_{st}\right]\right)$.

The SAST algorithm explores the whole state space $\left|\mathcal{SC}^{CQI,CT}\right|$ of center sets for both decisions by using the probability function of moving from the center set space $\mathcal{SC}_{tk}^{CQI,CT}$ to the newest one $\mathcal{SC}_{tk}^{'CQI,CT}$, the non-better solution acceptance probability Pr_{Ac} and the transition probability from Swap to Lloyd $Pr_{S\to L}$. For each of these probabilities, a local distortion parameter is proposed [196], such as: consecutive RDL $\left(RDL_C\right)$ for the transition probability $Pr_{S\to L}$ from Swap to Lloyd and accumulated RDL $\left(RDL_A\right)$ for non-better solution acceptance probability Pr_{Ac}. Both of parameters are calculated based on Eq. 4.20 in which $D_{INIT}^{CRDL} = D\left(\mathcal{SC}_{tk,t_{st}-1}^{CQI,CT}\right)$ is the initial distortion for consecutive RDL and $D_{INIT}^{ARDL} = D\left(\mathcal{SC}_{tk,Best}^{CQI,CT}\right)$ is the initial distortion for the accumulated RDL. Based on the stochastic tunneling methodology, the probabilistic functions take the forms of Eq. 4.23 and Eq. 4.24:

$$Pr_{Ac}\left[\left(\mathcal{SC}_{Best}^{CQI,CT}\right)\to\left(\mathcal{SC}_{tk,t_{st}}^{CQI,CT}\right)\right] = exp\left(-\frac{F_{ST}\left[RDL_A\left(\mathcal{SC}_{tk,t_{st}}^{CQI,CT}\right)\right] - F_{ST}\left[RDL_A\left(\mathcal{SC}_{Best}^{CQI,CT}\right)\right]}{T_{Ac}}\right) \tag{4.23}$$

$$Pr_{S\to L}\left[\left(\mathcal{SC}_{tk,t_{st}}^{CQI,CT}\right)\to\left(\mathcal{SC}_{tk,t_{st}+1}^{'CQI,CT}\right)\right] = exp\left(-\frac{F_{ST}\left[RDL_C\left(\mathcal{SC}_{tk,t_{st}+1}^{'CQI,CT}\right)\right] - F_{ST}\left[RDL_C\left(\mathcal{SC}_{tk,t_{st}}^{CQI,CT}\right)\right]}{T_{S\to L}}\right) \tag{4.24}$$

where the state of center sets transition $\left(\mathcal{SC}_{tk,t_{st}}^{CQI,CT}\right)\to\left(\mathcal{SC}_{tk,t_{st}+1}^{'CQI,CT}\right)$ denotes the Monte Carlo step between two states of center sets, $T_{Ac} = T_{S\to L} = T_{Ac,S\to L}$ are the current temperatures for both probability decisions, and the stochastic tunneling function [192] is $F_{ST}\left(X\right) = 1 - \exp\left[-\left(RDL_{A(C)}\left(X\right) - RDL_{A(C)}^{Best}\right)/\gamma_{ST}\right]$ with the

stochastic tunneling parameter γ_{ST}, RDL_A^{Best} is the accumulated RDL for the best set of centers $\mathcal{SC}_{Best}^{CQI,CT}$ in which the average distortion is minimized and RDL_C^{Best} is the best consecutive RDL for a given center set $\mathcal{SC}_{tk,t_{st}}^{CQI,CT}$. Equations 4.23 and 4.24 perform the tunneling procedure through regions with local minima solutions.

The most important parameters in the SAST heuristic are the tunneling parameter γ_{ST} and the *hot* temperature $T_{Ac,S\to L}$. The annealing schedule aims to work with two loops: in the outer loop the hot temperature $T_{Ac,S\to L}$ is decreased based on the cooling schedule at each run epoch and the inner loop, performs the decision making at each stage, based on the same hot temperature level. Two main problems arise in these situations: the initial hot temperature setting and the cooling schedule function. The initial value of the hot temperature can be determined based on accumulated or consecutive $RDL_{A(C)}$ for a given number of stages if other initial probability acceptance does not alter the accumulated or consecutive Relative Distortion Loss $RDL_{A(C)}$. In other words, the initial hot temperature $T_{Ac,S\to L}$ is calculated for a number of N_{Stages}^{Temp} stages based on Eq. 4.25 [185], where $\Delta RDL_{A,C}(X) = RDL_{A,C}(X) - RDL_{A,C}^{Best}$ is the difference between two consecutive accumulated or consecutive RDLs. Then, the hot temperature is decreased based on the temperature reduction factor (R_T) at each predefined number of stages as shown by Eq. 4.26 [185]:

$$T_{Ac,S\to L}^{Initial} = exp\left[-\frac{\sum_{g=0}^{N_{Stages}^{Temp}}\left[\Delta RDL_{A,C}\left(\mathcal{SC}_{tk,t_{st}+g+1}^{'CQI,CT}\right) - \Delta RDL_{A,C}\left(\mathcal{SC}_{tk,t_{st}+g}^{CQI,CT}\right)\right]}{\gamma_{ST}\cdot N_{Stages}^{Temp}}\right] \tag{4.25}$$

$$T_{Ac,S\to L}\left[t_{run}\right] = T_{Ac,S\to L}\left[t_{run}-1\right]\cdot R_T \tag{4.26}$$

Based on the algorithm description, it can be concluded that (P_{SAST}) is similar in some senses to DSR-SMOO and DSR-CMOO problems with the amendment that for the CQI center determination, the uncontrollable set of data points (observations) has a finite size, whereas in the stochastic scheduling

problems, the searching space is practically infinite. The details of the proposed SAST meta-heuristic method for k-means clustering are presented in Algorithm 4.4 and the dynamic Swap-Lloyd reasoning is highlighted by Algorithm 4.3.

Algorithm 4.3 The Dynamic Lloyd-Swap Heuristic Based on Simulated Annealing with Stochastic Tunneling

Require: run^{New} : boolean variables which denote if there is a new run epoch
$m\_swap[t_{ts}]$: boolean variable which requires (or not) a swap stage

1. **while** $\left(t_{st} \le N_{st}^{max}\right)$
2. **start** stage t_{st}
3. **if** $\left(run^{New} = true\right)$ $run^{New} = false$, $t_{run}^{st} = 0$
4. **end if**
5. **if** $\left(m\_swap[t_{st}] = true\right)$
6. **do Swap (see Algorithm 4.2) – Lines: 8, 9**
7. **else if**
8. **do Lloyd (see Algorithm 4.1) – Lines: 11, 12, 13, 14**
9. **end if**
10. $t_{run}^{st} = t_{run}^{st} + 1$
11. **if** $\left[\left(t_{run}^{st} \le N_{st/run}^{max}\right) \&\& \left(RDL_C[t_{st}] > RDL_C^{Min}\right)\right]$
12. **if** $\left(m\_swap[t_{st}] = true\right)$
13. **if** $\left(SASTAcceptance\left(RDL_C[t_{st}], t_{st}\right) = true\right)$
14. $m\_swap[t_{st}] = false$, $run^{New} = false$
15. **end if**
16. **end if**
17. **else if** $\left(t_{run}^{st} = N_{st/run}^{max}\right)$
18. **Decrease** $T_{Ac,S\to L}[t_{run}]$ **based on 4.26**
19. $m\_swap[t_{st}] = true$, $run^{New} = true$
20. **end if**
21. $\mathcal{S}_{Save}^{CQI,CT} = \mathcal{S}_{t_{st}}^{CQI,CT}$
22. **if** $\left(\overline{D}[t_{st}] < \overline{D}_{Best}\right)$
23. $\mathcal{S}_{Best}^{CQI,CT} = \mathcal{S}_{Save}^{CQI,CT}$
24. **else if** $\left(SASTAcceptance\left(RDL_A[t_{st}], t_{st}\right) = true\right)$
25. $\mathcal{S}_{Best}^{CQI,CT} = \mathcal{S}_{Save}^{CQI,CT}$
26. **else**
27. $\mathcal{S}_{t_{st}}^{CQI,CT} = \mathcal{S}_{Save}^{CQI,CT}$
28. **end if**
29. **end** stage t_{st}
30. **end while**

Algorithm 4.4 Simulated Annealing with Stochastic Tunneling

$bool \quad SASTcceptance\left(RDL_{A(C)}\left[t_{st}\right], t_{st}\right)$

1. **if** $\left(RDL^{Best}_{A(C)} < RDL_{A(C)}\left(\mathcal{SC}^{CQI,CT}_{tk,t_{st}}\right)\right)$
2. $\quad RDL^{Best}_{A(C)} = RDL_{A(C)}\left(\mathcal{SC}^{CQI,CT}_{tk,t_{st}}\right)$
3. **end if**
4. $\Delta RDL_{A(C)}\left[t_{st}-1\right] = RDL_{A(C)}\left[t_{st}-1\right] - RDL^{Best}_{A(C)}$
5. $\Delta RDL_{A(C)}\left[t_{st}\right] = RDL_{A(C)}\left[t_{st}\right] - RDL^{Best}_{A(C)}$
6. $F_{ST}\left[t_{st}-1\right] = 1 - exp\left(\Delta RDL_{A(C)}\left[t_{st}-1\right] / \gamma_{ST}\right)$
7. $F_{ST}\left[t_{st}\right] = 1 - exp\left(\Delta RDL_{A(C)}\left[t_{st}\right] / \gamma_{ST}\right)$
8. **if** $\left(t_{st} \le N^{Temp}_{Stages}\right)$
9. $\sum RDL_{A(C)}\left[t_{st}, t_{st}-1\right] = \sum RDL_{A(C)}\left[t_{st}-1, t_{st}-2\right] + \left(RDL_{A(C)}\left[t_{st}\right] - RDL_{A(C)}\left[t_{st}-1\right]\right)$
10. $\quad$ **if** $\left(t_{st} = N^{Temp}_{Stages}\right)$
11. $\quad\quad$ **Calculate** $T^{Initial}_{Ac,S\to L}$ **based on Eq. 4.25**
12. $\quad$ **end if**
13. $\quad Pr_{Ac,S\to L} = Pr_{Init}$
14. **else**
15. $\quad$ **Calculate** $Pr_{Ac,S\to L}$ **based on Eq. 4.23 and 4.24**
16. **end if**
17. **Return** $Pr_{Ac,S\to L} > P_{Rand}$

4.5.2 Supervised Learning in CQI Classification

In the supervised learning step, it is assumed that the set of preprocessed CQI data centers $\mathcal{S}^{CQI,CT} = \mathcal{SC}^{*CQI,CT}_{tk} = \left\{\mathcal{S}^{CQI,CT}_{k}, k = 1,..,N_{CT}\right\}, \forall tk = 1,...,\left|\mathcal{SC}^{CQI,CT}\right|$ is already known according to the unsupervised learning step. A very important characteristic refers to the possibility of connecting the supervised learning step directly to the CQI cycle module. In other words, the classification stage can be trained based on the observations received from the LTE network environment in what is called *online training* of the preprocessed CQI reports. The supervised learning stage should classify each new entry in one of the predefined cluster obtained in the previous section. The name of supervised stands with the idea that for some input pairs, the output is labeled or is already known, and the generalization procedure for other observations is achieved based on the known

patterns. For the purpose of the CQI report classification, a tricky method is used in order to determine the patterns. For example, the output value can be determined by calculating the Euclidian distance between a given input and each center point. Then, the input observation is associated with the nearest k-means center, and the center index is transformed from decimal to binary representation.

As mentioned in the early stage of the section, several problems can appear during the classification procedure. The main target is to minimize the error between the desired and the obtained output sets. The classification process aims to minimize the error for each given set of CQI observations. In practice, this aspect is impossible to be achieved due to the modality of how the CQI inputs are provided to the classifier. When connected to the CQI LTE cycle module with the Zheng's model fading type (see Appendix B), the classifier will fall in the local minima. By providing for a longer number of epochs the same (or appropriate) set of CQI observations (Fig. C.4.b and Fig. C.5.b from Appendix C), when the preprocessed CQI distribution changes, the newest entries are classified based on the local minimum detected so far. This is the main reason why it was decided to introduce the LTE CQI cycle section in order to highlight the necessity of the Jakes model with very fast fading for the classifier input. By providing a dynamic distribution of the preprocessed CQI reports (Fig. C.4.a and Fig. C.5.a from Appendix C), the classifier is able to offer a better generalization of the predicted outputs and an enhanced classification performance. But even with the Jakes fading model, the local minima avoidance is not totally solved.

The *golden part* of this section divides the set of CQI observations into two data sets: ***validation and training sets***. The validation set is basically the collected set of the preprocessed CQI reports without duplicates obtained over a long time of the collection process for each LTE bandwidth and for different preprocessing scheme settings. The observations from the validation step can be applied through the training procedures together with the training sets in order to find the global minima. The problem that arises in this case has to deal with the decision when the validation observation should be applied. The same ***SA algorithm with Stochastic Tunneling*** is proposed at this stage to switch from training to validation sets and vice-versa by monitoring the epoch errors for both

sets. It is important to notice that the number of epochs within 1 TTI depends on the number of CQI reports and implicitly on the number of active users.

The RBFNN with *backward propagation* learning [165], [166] is used for the classification of the preprocessed CQI inputs. The proposed architecture respects the following steps in training and updating the neural network weights:

1. ***Training and validation***: Based on the epoch error, the SAST method decides if the training observation provided by the CQI cycle module should be used as an observation for the RBFNN structure or a new sample from the validation set should be randomly chosen.
2. ***The feed-forward computation*** in which the input is passed through the RBFNN layers and a new predicted output is obtained.
3. ***The error back-propagation and the weight corrections:*** The output error is calculated based on the obtained output value under the provided pattern and then ***back-propagated*** to the input layers. The updating process of the RBFNN weights is based on the *gradient descent principle* [165].

The input pattern set is equivalent to the best set of preprocessed CQI centers $\mathcal{S}^{CQI,CT}=\left\{\mathcal{S}_k^{CQI,CT},k=1,..,N_{CT}\right\}$ being obtained in the clustering process. For simplicity, let us define the TTI $t=\left|\mathcal{U}_t\right|\cdot t_{eph}$, where t_{eph} is the RBFNN epoch time instant. This approach is very advantageous due to the fact that the learning speed is greater than that in other cases of RL algorithms, mainly due to a higher number of observations which can be received in one TTI.

For the preprocessed CQI classification purpose, the output dimension of the RBFNN structure is $N_O^{RBF}=ln\left(N_{CT}\right)$ which means that for 64 centers, the output dimension is 6. The output pattern set is determined in two steps:

1. The ordered set of centers $\mathcal{S}_{od}^{CQI,CT}$ is obtained by using the Euclidian distance from the obtained set of preprocessed CQI centers $\mathcal{S}^{CQI,CT}=\mathcal{SC}_{tk}^{*CQI,CT}$, $\forall tk=1,...,\left|\mathcal{SC}^{CQI,CT}\right|$ to the set of *support centers* $\mathcal{S}_{sup}^{CQI,CT}=\left\{MCQI_v\right\},v=1,..,N_{CQI}$ where $\sum_{v=1}^{N_{CQI}}MCQI_v=1$, where the top mass CQI scheme is $Top_{CQI}=1$ and the set size of the support centers is

$\left|\mathcal{S}_{sup}^{CQI,CT}\right| = 15$. When calculating the Euclidian distance $\left\|\mathcal{S}^{CQI,CT} - \mathcal{S}_{sup}^{CQI,CT}\right\|$, the original centers set becomes the ordered set of preprocessed CQI centers $\mathcal{S}_{od}^{CQI,CT} = \left\{\mathcal{S}_{k_{od}}^{CQI,CT}, k_{od} = 1,..,N_{CT}\right\}$, meaning that when $k_{od} = 1$, the center set denotes the worst channel quality from the whole set, and when $k_{od} = N_{CT}$, the center set represents those users with favorable channel feedbacks being reported at each TTI.

2. Once the obtained center set is ordered, the output RBFNN pattern set $\mathcal{S}_{O,Patt}^{CQI,C} = \left\{\mathcal{S}_{op}^{CQI,C}, op = 1,..,N_O^{RBF}\right\} \in \{0,1\}$ represents the binary translation from the corresponding center index k_{od}.

For each new observation of the preprocessed CQI vector, the closest center index from the ordered set of preprocessed CQI centers $\mathcal{S}_{od}^{CQI,CT}$ is determined and the output RBFNN pattern $\mathcal{S}_{O,Patt}^{CQI,C}$ is the binary version of the closest center index. In this case, the output RBFNN pattern $\mathcal{S}_{O,Patt}^{CQI,C}$ represents the desired output. Then, the observation is feed-forwarded through the RBFNN structure and the predicted output is determined in terms of the classified preprocessed CQI state space $\mathcal{S}_{O}^{CQI,C}$ at the RBFNN output layer. The instantaneous quadratic error between the predicted and the pattern outputs which is calculated at each epoch is revealed by Eq. 4.27 [166]:

$$E\left(\mathcal{S}_{O,t_{eph}}^{CQI,C}, \mathcal{S}_{O,Patt}^{CQI,C}\right) = 1/2 \cdot \sum_{o=1}^{N_O^{RBF}} \left\|\mathcal{S}_{o,t_{eph}}^{CQI,C} - \mathcal{S}_{o,Patt}^{CQI,C}\right\|^2 \tag{4.27}$$

where t_{eph} represents the time instant when the RBFNN receives a new preprocessed CQI observation. Therefore, the objective of the back-propagation structure is to minimize the mean or average error $\overline{E}\left(\mathcal{S}_{O,t_{eph}}^{CQI,C}, \mathcal{S}_{O,Patt}^{CQI,C}\right)$ between the classified preprocessed CQI state at the RBFNN output layer and its corresponding pattern, as shown by Eq. 4.28:

$$min\left[1/N_{eph} \cdot \sum_{ep=1}^{N_{eph}} E\left(\mathcal{S}_{O,ep}^{CQI,C}, \mathcal{S}_{O,Patt}^{CQI,C}\right)\right] \tag{4.28}$$

where N_{eph} is the number of epochs involved in the RBFNN training procedure.

4.5.2.1 Training and Validation

The local minima problem can be avoided by using the validation set which is considered to be different from the training one. Consequently, both types of errors for training and validation sets should be monitored such as E_{Train}^{RBFNN} (the RBFNN error when the inputs are provided from the training set) and E_{Val}^{RBFNN} (the RBFNN error when the inputs are provided from the validation set). The mean error for both sets can be calculated based on Eq. 4.28 averaged over N_{eph}^{Train} number of training epochs or over N_{eph}^{Val} number of validation epochs. The mean training error $\overline{E}_{Train}^{RBFNN}$ aims to decrease over the time when the training stage is performed. On the other hand, based on the frequency of using the validation set, the mean error for the validation set $\overline{E}_{Val}^{RBFNN}$ decreases for some number of epochs and then starts to increase gradually. This effect is called over-fitting and represents a problematic issue in the prediction problems. Fortunately, based on the experimental results from Section 4.7 and based on the proper parameterization of the RBFNN structure, the preprocessed CQI classification *is not the subject of the over-fitting symptom*.

For the local minima problem, the consecutive epoch error loss $CEL\left[t_{eph}\right]$ for both training and validation sets is monitored according to Eq. 4.20 and reloaded in Eq. 4.29 for a comprehensive representation:

$$CEL\left[t_{eph}\right]=\frac{E_{Train(Val)}^{RBFNN}\left[t_{eph}-1\right]-E_{Train(Val)}^{RBFNN}\left[t_{eph}\right]}{E_{Train(Val)}^{RBFNN}\left[t_{eph}-1\right]} \tag{4.29}$$

When the consecutive epoch error loss is $CEL\left[t_{eph}\right]>0$, the newest observation fits better under the learned RBFNN surface. The main problem is the prediction on which observation type (training or validation) should be chosen in order to find a better minimum solution when compared with the existing one. In this case, the same principle of the SAST scheduler is applied. The SAST methodology aims to select, based on the consecutive epoch error loss $CEL\left[t_{eph}\right]$, which type of observation should be applied to the RBFNN structure. The initial hot

temperature is calculated based on Equation 4.30 and the temperature reduction factor is similar to Equation 4.26:

$$T_{Train->Val}^{Initial} = exp\left[-\frac{\sum_{g=0}^{N_{Epochs}^{Temp}} \left[\Delta CEL\left(\mathcal{S}_{O,t_{eph}+g+1}^{CQI,C} \right) - \Delta CEL\left(\mathcal{S}_{O,t_{eph}+g+1}^{CQI,C} \right) \right]}{\gamma_{ST} \cdot N_{Epochs}^{Temp}} \right] \tag{4.30}$$

The probability of acceptance for a validation observation keeps the same form as indicated in Eq. 4.24. Once the observation type is decided to be applied at epoch t_{eph}, the selected preprocessed CQI input is feed-forwarded through the RBFNN structure in order to get the classified preprocessed CQI output.

4.5.2.2 The Feed-Forward Computation

The perfect interpolation of Equation 4.13 is practically impossible because of two main aspects:

- the finite preprocessed CQI collection set size, and
- the non-convexity property of the k-means clustering problem: the set of optimal centers are very difficult to be find.

Therefore, the function approximation is used in the CQI state space aggregation under the form of the RBFNN structure. When the set of observations for a given epoch t_{eph} is approximated through the experiences learned so far, the procedure is entitled the feed-forward propagation. Figure 4.5 shows the RBFNN usage in the CQI classification procedure, and Figure 4.6 brings a clearer explanation about the RBFNN feed-forward computation which is executed for each layer and for each node based on each CQI observation provided at each epoch.

The input layer represented by the preprocessed CQI state space $\mathcal{S}_{i,t}^{CQI,P,TM}$ with top or majority mass modes is directly passed to all hidden nodes and the activation function $\varphi_h = \varphi_h\left(\mathcal{S}_{i,h,t_{eph}}^{CQI,P,TM}, \mathcal{S}_h^{CQI,CT} \right)$ is calculated. It is assumed that the RBFNN activation function for the input layer is linear and the activation function for the hidden layer takes the Gaussian form such that [166], [167]:

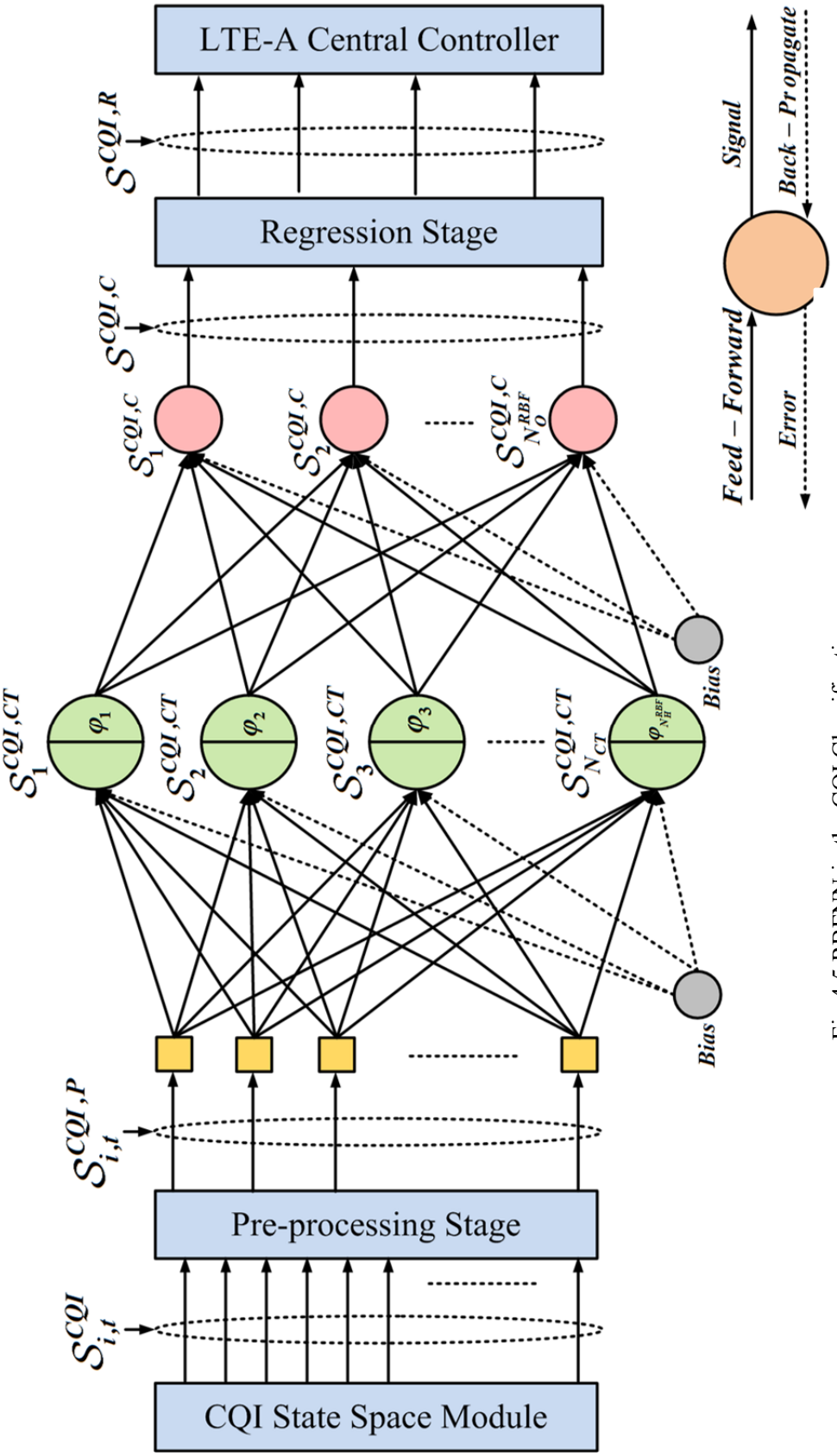


Fig.4.5 RBFNN in the CQI Classification

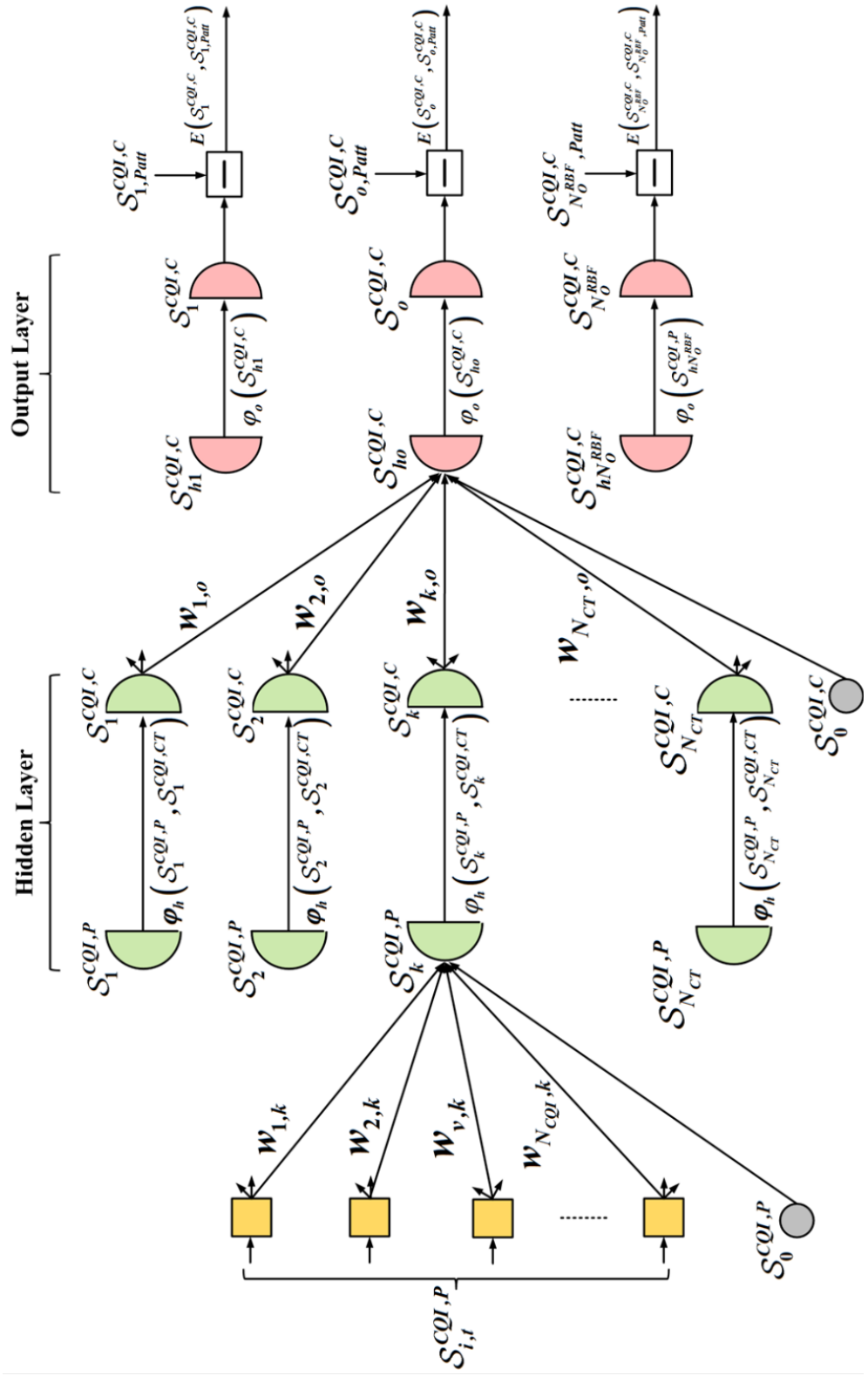


Fig.4.6 The RBFNN Feed-Forward Computation

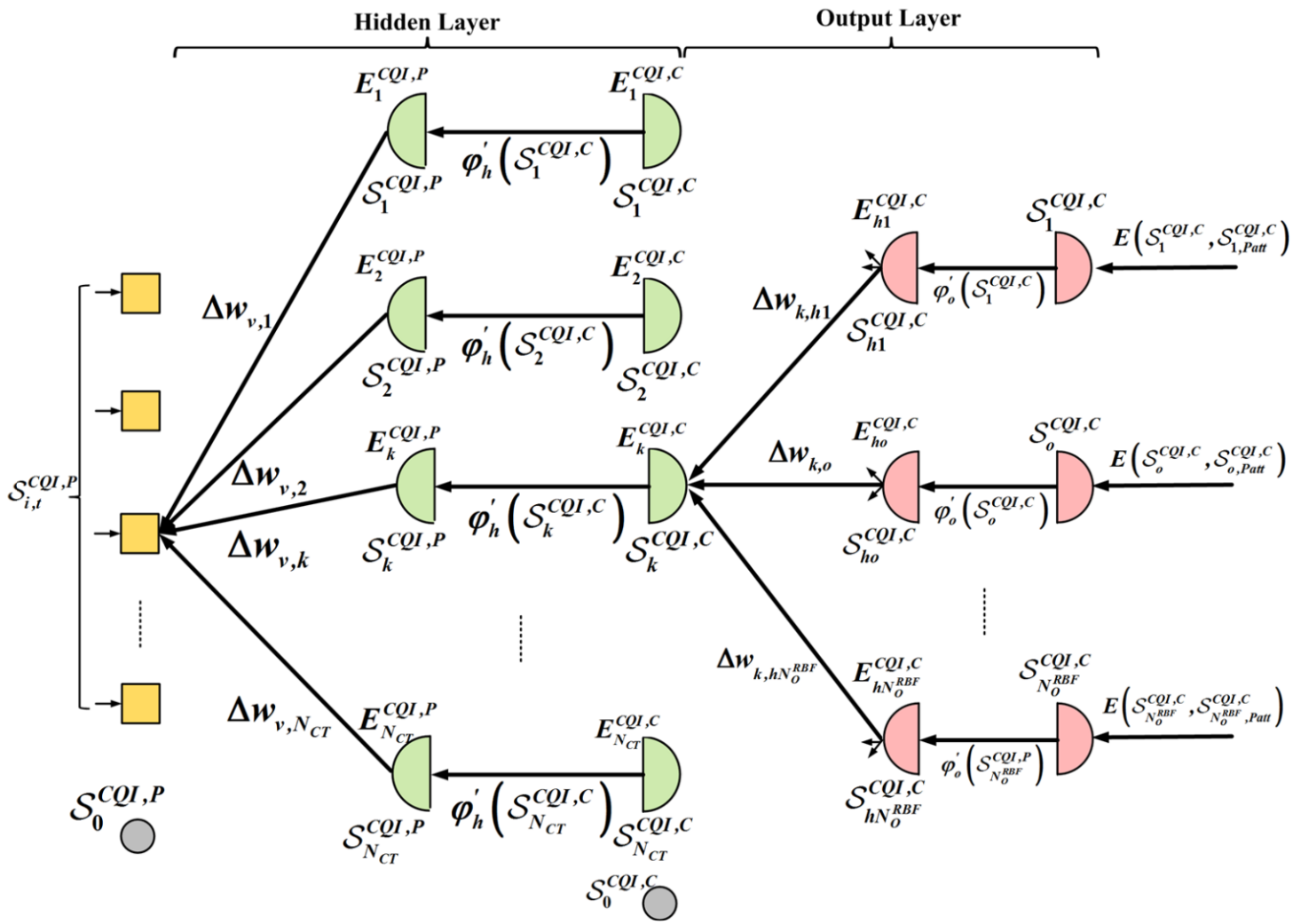


Fig.4.7 The RBFNN Error Backward Propagation

$$\varphi_h\left(\mathcal{S}_{i,h,t_{eph}}^{CQI,P,TM},\mathcal{S}_k^{CQI,CT}\right)=exp\left(-\sigma_{RBF}\cdot\left\|\mathcal{S}_{i,h,t_{eph}}^{CQI,P,TM}-\mathcal{S}_k^{CQI,CT}\right\|\right) \quad (4.31)$$

where σ_{RBF} is the Gaussian weight which plays a crucial role in the RBFNN mean squared error minimization. In order to reduce the computational complexity of hidden nodes, the Gaussian weights are obtained through extensive simulation results for different sets of k-means centers and top mass schemes which can be used by the LTE-A central controller. Based on the Equations 4.13 and 4.14 and based on Figure 4.5 and Figure 4.6, the output node of the RBFNN network can be computed in the following manner:

$$\mathcal{S}_{o,i}^{CQI,C}=\varphi_o\left(\sum_{k=1}^{N_{CT}}w_{k,o}\cdot\varphi_k\left(\mathcal{S}_{i,h,t_{eph}}^{CQI,P,TM},\mathcal{S}_k^{CQI,CT}\right)\right) \quad (4.32)$$

where φ_o is the activation function for the output layer. As mentioned earlier, for interpolation reasons, the number of hidden nodes is equal to the number of k-means SAST centers. By computing in a similar way to other outputs nodes, the output RBFNN classified state space $\mathcal{S}_{O,i}^{CQI,C}$ for user $i\in\mathcal{U}_t$ is obtained. As shown by Figure 4.6, the instantaneous RBFNN error for each node is obtained based on the difference between the predicted and the pattern outputs. The next steps in the RBFNN training are the error backward propagation and the necessary procedures for the RBFNN weight corrections which are performed epoch-by-epoch.

4.5.2.3 The Backward Error Computation

The backward propagation acts in the opposite direction when compared with the feed-forward propagation and comprises two functions: weight corrections and the error backward propagation for each node and for each layer. Without going through deep details about the backward propagation principles, this study strictly focusses on the CQI classification purpose for different collections of the preprocessed CQI vectors with different bandwidths. More details about the error propagation techniques are detailed in [165], [166], [167].

Figure 4.7 shows the backward propagation architecture which performs the error propagation stage that is implied in the RBFNN training procedure.

Basically, the corrections of the RBFNN weights are achieved based on the gradient descent training which permits to adjust them in the opposite direction of the gradient objective function for each node [165]. In other words, at each node level, the derivative activation function is performed according to the input and the output values for each node. The back-propagated error $E_{ho}^{CQI,C}$ of the classified preprocessed CQI observation for the *ho* node at the input of the output layer is calculated according to the following equation:

$$E_{ho}^{CQI,C} = \varphi_o^{'}\left(\mathcal{S}_o^{CQI,C}\right) \cdot E\left(\mathcal{S}_o^{CQI,C}, \mathcal{S}_{o,Patt}^{CQI,C}\right) \tag{4.33}$$

where $\varphi_o^{'}$ represents the output derivative function. The weight correction is applied according to the back-propagated error of $E_{ho}^{CQI,C}$ and the output value of the k^{th} hidden node such that:

$$\Delta w_{k,o} = \eta_{RBF} \cdot E_{ho}^{CQI,C}\left[t_{eph}\right] \cdot \mathcal{S}_k^{CQI,C}\left[t_{eph} - 1\right] \tag{4.34}$$

where η_{RBF} is the learning parameter [165]. The learning parameter denotes the learning speed or the correction step length. In this study, the optimal pairs of learning parameters and Gaussian weights are obtained through extensive simulation results for different types of preprocessing configurations. Therefore, the RBFNN weight value is updated by simply adding the correction such that $w_{k,o}\left[t_{eph}\right] = w_{k,o}\left[t_{eph} - 1\right] + \Delta w_{k,o}\left[t_{eph}\right]$ for each hidden node *k* and for each output node *o*. The back-propagated error for the k^{th} output hidden unit is given by:

$$E_k^{CQI,C} = \sum_{o=1}^{N_O^{RBF}} w_{k,o} \cdot E_{ho}^{CQI,C} \tag{4.35}$$

The error for input hidden unit becomes: $E_k^{CQI,P} = \varphi_h^{'}\left(\mathcal{S}_k^{CQI,C}\right) \cdot E_k^{CQI,C}$. Following the same principle of Eq. 4.34, the weight correction for the input layer becomes $\Delta w_{v,k} = E_k^{CQI,P} \cdot \mathcal{S}_v^{CQI,P}$. The learning parameter is not included since it was introduced in Eq. 4.34. The backward propagation principle is performed in conjunction with the feed-forward step until the output error falls under the desired threshold after a given number of training epochs.

4.6 Regression Stage in CQI Aggregation

For each 15-dimensional preprocessed CQI observation, the RBFNN provides an output vector of $D\left[\mathcal{S}_{O,i}^{CQI,C}\right]=ln\left(N_{CT}\right)$ dimension for each UE $i\in\mathcal{U}_t$. The RBFNN structure can provide the cluster index for each given preprocessed CQI inputs if an additional processing unit will be considered. In this sense, a threshold for each RBFNN output node should be defined in order to discretize the continuous output domain. In many classification problems, the output activation function is considered to be sigmoid or tangent hyberbolic [201], [202]. For the current purpose, the output function is considered to be tangent hyperbolic and it is defined by Eq. 4.36:

$$\varphi_o:\mathbb{R}\rightarrow[-1,1] \qquad \varphi_o\left(\mathcal{S}_{ho}^{CQI,C}\right)=tanh\left(\mathcal{S}_{ho}^{CQI,C}\right) \tag{4.36}$$

where the derivative is $\varphi_o^{'}\left(\mathcal{S}_{o}^{CQI,C}\right)=1-\left(\mathcal{S}_{o}^{CQI,C}\right)^2$. The output threshold is decided to classify each output node in two discrete classes: 0 and 1. Let us define the RBFNN output threshold as $\mathcal{S}_{o,TH}^{CQI,C}$. For the preprocessed CQI classification, the RBFNN output threshold is fixed to $\mathcal{S}_{o,TH}^{CQI,C}=0$. Then, the discretized version of the output node *o* can be decided based on Eq. 4.37:

$$\mathcal{S}_{o,i,D}^{CQI,C}\left[t_{eph}\right]=\begin{cases}0,\ if\ \mathcal{S}_{o,i}^{CQI,C}\left[t_{eph}\right]<\mathcal{S}_{o,TH}^{CQI,C}\\1,\ if\ \mathcal{S}_{o,i}^{CQI,C}\left[t_{eph}\right]\geq\mathcal{S}_{o,TH}^{CQI,C}\end{cases} \tag{4.37}$$

Hence, the discretized output RBFNN state space $\mathcal{S}_{O,i,D}^{CQI,C}$ keeps a similar dimension to the classified preprocessed CQI state at the RBFNN output layer $\mathcal{S}_{O,i}^{CQI,C}\left[t_{eph}\right]=\bigcup_{o=1}^{N_O^{RBF}}\mathcal{S}_{o,i}^{CQI,C}\left[t_{eph}\right]$ for each user $i\in\mathcal{U}_t$. By simply transforming the discretized binary vector $\mathcal{S}_{O,i,D}^{CQI,C}\left[t_{eph}\right]=\bigcup_{o=1}^{N_O^{RBF}}\mathcal{S}_{o,i,D}^{CQI,C}\left[t_{eph}\right]$ to decimal, the obtained value from the classified preprocessed CQI state at the RBFNN output layer $\mathcal{S}_{O,i,D}^{CQI,C}$ is the center index which represents the CQI feedback for each user.

At each TTI, all active users report the preprocessed CQIs to the RBFNN structure. The training procedure of the RBFNN module is achieved sequentially,

and the RBFNN classified output space is obtained at each TTI. From the LTE scheduler controller point of view, the general information about the channel quality is needed for a proper selection of the scheduling rules. Precisely, the controller should have the exact information about how many users belong to the considered clusters. This fact implies in the first instance the discretized classified preprocessed CQI state space $\mathcal{S}_{O,i,D}^{CQI,C}$ expansion to $\mathcal{S}_{O}^{CQI,C}=\bigcup_{i=1}^{|\mathcal{U}_t|}\mathcal{S}_{O,i,D}^{CQI,C}$, where the dimension of the overall classified preprocessed CQI state at the RBFNN output layer is $D\left[\mathcal{S}_{O}^{CQI,C}\right]=|\mathcal{U}_t|\times log\left(N_{CT}\right)$. The obtained state space contains the center index for each reported feedback at each TTI *t*. Then, the classified state space $\mathcal{S}_{t}^{CQI,CL}$ at TTI *t* can be obtained by counting the number of preprocessed CQI reports which belongs to cluster *k* at TTI *t,* such that:

$$\mathcal{S}_{t}^{CQI,CL}=\bigcup_{k=1}^{N_{CT}}\left\{\frac{N_U^k[t]}{|\mathcal{U}_t|}\right\} \tag{4.38}$$

where $N_U^k[t]$ is the number of users belonging to class *k* and the dimension of the obtained state space is $D\left[\mathcal{S}_{t}^{CQI,CL}\right]=N_{CT}$. Based on the classified state space, the regression processing unit aims to extract the most relevant information at each TTI which is considered crucial for the LTE scheduler controller, such as:

1. **<u>The number of active clusters</u>**. The active cluster contains at least one active user. Let us define N_{CL}^{A} as the number of active clusters which can be expressed as follows:

$$\mathcal{S}_{1,t}^{CQI,R}=N_{CL}^{A}[t]=\sum_{k=1}^{N_{CT}}n_s^k[t],\ n_s^k[t]=\begin{cases}1,\ if\ N_U^k[t]\neq 0\\0,\ if\ N_U^k[t]=0\end{cases} \tag{4.39}$$

   If all active users belong to one class, then the $\mathcal{S}_{t}^{CQI,CL}$ state becomes the *classified support vector*.

2. **<u>The STD of the percentage of users belonging to different active clusters</u>**. For the DSR-SMOO problems focusing on the NGMN fairness requirement, the GPF scheduling rule parameterization can be learned based on the dispersion of the percentage of users across active clusters

which can be calculated according to Eq. 4.40:

$$\mathcal{S}_{2,t}^{CQI,R} = \sigma_{CL}[t] = \frac{1}{N_{CL}^{A}[t]} \cdot \sum_{k=1}^{N_{CT}} n_a^k[t] \cdot \left(\frac{N_U^k[t]}{|\mathcal{U}_t|} - 1 \right)^2, n_a^k[t] = \begin{cases} 1, \ if \ N_U^k[t] \neq 0 \\ 0, \ if \ N_U^k[t] = 0 \end{cases}$$

(4.40)

3. **<u>The closest support vector index</u>.** As mentioned earlier, the classified support vector can be defined as $\mathcal{S}_{k,Supp}^{CQI,CL} = \{n_k\}$, $n_k \in \{0,1\}$, where $\sum_{k=1}^{N_{CT}} n_k = 1$ and the classified support vector space size is $\left|\mathcal{S}_{Supp}^{CQI,CL}\right| = N_{CT}$. Thus, the closest support vector index can be defined as follows:

$$\mathcal{S}_{3,t}^{CQI,R} = ks[t] = \underset{k}{arg\,min}\left[\left\| \mathcal{S}_{k,Supp}^{CQI,CL} - \mathcal{S}_{t}^{CQI,CL} \right\| \right] \quad (4.41)$$

4. **<u>The Euclidian distance between the classified output state and the closest support vector</u>.** Based on the closest support vector $\mathcal{S}_{3,t}^{CQI,R}$, the distance from the classified preprocessed CQI state at TTI t $\mathcal{S}_{t}^{CQI,CL}$ to its closest class support vector is required in order to enhance the integrity of the regressed state space:

$$\mathcal{S}_{4,t}^{CQI,R} = d_{ks}[t] = \left\| \mathcal{S}_{ks[t],Supp}^{CQI,CL} - \mathcal{S}_{t}^{CQI,CL} \right\| \quad (4.42)$$

Based on Equations 4.39, 4.40, 4.41 and 4.42, the regressed CQI state space at TTI t is denoted by Eq. 4.43:

$$\mathcal{S}_{t}^{CQI,R} = \left\{ N_{CL}^{A}[t], \sigma_{CL}[t], ks[t], d_{ks}[t] \right\} \quad (4.43)$$

The obtained state space contains statistical elements of the classified state space which represents the uncontrollable input state space for the LTE-A scheduler controller. When multiple traffic types with different priorities are considered, each class has its own regressed preprocessed CQI state space. Then, the aggregation procedure for the uncontrollable CQI state space becomes more important due to its ability of reducing the overall CQI state space dimension from $N_P \times |\mathcal{U}_t| \times |\mathcal{B}|$ to $N_P \cdot \left|\mathcal{S}_{t}^{CQI,R}\right|$, where N_P is the number of traffic priorities. Precise details about the LTE controller architecture for the DSR-SMOO/CMOO problems with the multi-class traffic types are provided in Chapter 8.

4.7 Performance Evaluation of CQI Aggregation

The simulation scenario and the experimental results are conducted through five directions based on the operating modes exposed in Section 4.4. The modalities about the ways how the experiments are conducted and about the ways how the results are gathered are also presented in the following.

4.7.1 Simulation Scenario

***Mode P:*1-*C:*0-*R:*0** corresponds to the preprocessing stage in which the CQI report becomes a 15-dimensional vector and the bandwidth dependency is avoided. In order to compress the preprocessed CQI stage space, different reassignment modes are exposed in Appendix C. For these simulation results, it is decided to use three types of configurations such as Top3, Top4 and Top5 CQI Mass Modes for each LTE bandwidth from [152]. These configuration modes present the best results from the viewpoint of the aggregation performance as indicated in Table C.1 from Appendix C. The Algorithm C.1 which is proposed in Appendix C is coupled to the CQI aggregation module in order to reduce the preprocessed CQI state space by using different configuration modes.

***Mode P:*1-*C:*1-*R:*0** reflects the collection procedure of the preprocessed CQI reports under different top mass configurations for each system bandwidth. Once the preprocessing configuration modes are set, the idea is to couple the Algorithms C.1 and C2 from Appendix C to the LTE CQI reporting module from Appendix B and to run the entire structure until the termination condition given by the Equation C.5 is fulfilled. By using the termination condition for the CQI state space collection with different LTE bandwidth configurations, the obtained state space collection size $\left|\mathcal{S}_t^{CQI,P,TM}\right|$ takes different values. The set of parameters for the LTE CQI feedback module is presented in Table 4.3, where most of the parameter settings are imported from the 3GPP specifications [36]. A large number of user reports $\left|\mathcal{U}_t\right| = 1000$ is simulated in order to speed-up the CQI collection procedure. The user speed is 120kmph in order to provide as many different CQI observations as possible and to enhance the collection procedure.

Table 4.3 CQI Feedback Module Parameter Settings

Parameter Name	Parameter description
Downlink Transmission Power (P^{TX})	43dBm [36]
Multipath Fading Model	Jakes Model with 12 Multiple Paths [36]
Path Loss Model	Macro Cell Urban Area [36]
Penetration Loss	10 dB [36]
Shadowing Loss Mean and Deviation	$\mu = 0$, $\sigma = 8dB$ [36]
Noise Figure (F)	2.5 [36]
Noise Spectral Density	-174dBm [36]
Number of Cells	19 [36]
Frequency Reuse Factor	3 [36]
RB Bandwidth	180KHz [36]
CQI reporting mode	Periodic at each 1ms [36]
PUCCH Channel	Errorless
Number of CQI feedbacks at each TTI	1000
User Speed	120kmph
User Mobility Model	Random Direction
LTE Bandwidth	Configurations from Table 4.1
CQI Preprocessing Mode	Top3, Top4, Top5 CQI Mass Modes
SINR Level Quantization	AWGN channel (Appendix B)
Target BLER	10% (Fig. B.11.a, Appendix B)

Table 4.4 The Size of the Preprocessed Top CQI Sets for Different LTE Bandwidth Configurations (based on the simulation results)

LTE Bandwidths [152]	Preprocessed CQI State Space Size		
	Top3 Mass Mode	Top4 Mass Mode	Top5 Mass Mode
1.4 MHz	2373	5070	7294
3 MHz	8749	38685	74433
5 MHz	13016	65691	117214
10 MHz	18531	90584	125470
15 MHz	24268	105755	162936
20 MHz	33596	144179	206473

From the same reasons, the CQI reporting mode is periodic at each TTI without delay and the PUCCH channel is considered to be errorless. The simulation scenario considers a number of 19 cells, but the collection procedure is launched only in the central cell and the rest of cells provide the interference model which is exposed in Appendix B, where the frequency reuse factor is 3 and the mobility model is random direction [68]. This is to increase the chances of getting new preprocessed CQI observations at each TTI. For CQI quantization reasons, the target BLER is set to 10% (see Appendix B). The obtained set sizes of the top mass preprocessed CQI observations for each LTE bandwidth configuration are

shown in Table 4.4. The forgetting factor β_{TM} from Equation C.5 is fixed to $\beta_{TM} = 0.005$, and when $\left|\overline{\mathcal{S}_{New,t}^{CQI,P,TM}}\right| > 0.99$, the collection procedure is stopped. The entire collection procedure was performed for more than 3 weeks of simulations by using the LTE-A-Scheduler simulator on 18 different machines. As mentioned earlier, the collections of the preprocessed CQI observations for each bandwidth are used in two ways: first, the k-means clustering algorithms are performed and the sets of the preprocessed CQI centers are obtained, and second, the collections represent the validation sets for each LTE bandwidth which are used in the RBFNN training stage.

***Mode P:*1-*C:*2-*R:*0** applies the proposed SAST clustering algorithm to the collected sets of preprocessed CQI observations. The clustering approach works autonomous only based on the observations received from the data base and the LTE CQI feedback module is disconnected. In order to highlight the advantages of the proposed clustering algorithm, two types of simulations are presented. In the first instance, the impact of the static number of centers is studied based only on the evolution of the average distortion stage-by-stage (Sub-section 4.7.2 and Appendix D). In the second instance, the impact of the number of centers is studied for each obtained collection (see Sub-section 4.7.3 and Appendix E). The hybrid SAST k-means clustering performance is compared against other existing methods in the literature such as hybrid-SA or hybrid-EZ schemes [185].

***Mode P:*1-*C:*3-*R:*0** trains the RBFNN structure based on the set of centers obtained in the previous working mode. The RBFNN structure considers only the set of centers with the best average distortion. Therefore, the procedure of the RBFNN weight corrections and the classification of the preprocessed CQI observations are performed based on the selected set of k-means centers and based on the selected observations (from the training or validation sets based on the SAST schedule). As mentioned, the validation set is in fact the collection set and the training set is provided when the preprocessing stage and RBFNN structure are connected to the LTE CQI feedback module. The learning parameter and the Gaussian weights are the most important elements in the RBFNN training process. When these parameters are not optimized, the average validation and training

errors are not able to decrease under a certain level as shown in Sub-section 4.7.4. In this sense, the optimization of these parameters is performed in Section 4.7.5 in order to find the best learning and Gaussian parameters which minimize the mean validation and training errors. Finally, the impact of optimized parameters in the mean output error evolution based on the number of time epochs is studied in Sub-section 4.7.6. If the RBFNN average error for the validation set continues to decrease uniformly, the training RBFNN algorithm is stopped, and the weights are saved when the decreasing rate for the total average RBFNN error becomes relatively constant for a given number of time epochs.

***Mode P:*1-*C:*4-*R:*0** evaluates the trained RBFNN structure from the previous working mode under different number of k-means centers as shown in Sub-section 4.7.7. Therefore, the working mode ***P:*1-*C:*4-*R:*1** can be performed in order to obtain the regressed CQI state space to be applied to the LTE-A scheduler controller for the scheduling policies adaptation and refinement.

4.7.2 Static Number of K-Means Centers

In the first instance, the iterated Lloyd, Swap and hybrid-SAST algorithms are performed for different sets of centers with different sizes reported to the stage number evolution. The performances of above mentioned algorithms are compared against the existing methods proposed in [185], such as hybrid-SA and hybrid-EZ. The latter one uses a combination of Lloyd and swaps in such a way that at the beginning of each running stage, the swap algorithm is applied only once and then, the hybrid-EZ approach continues with the Lloyd algorithm for the rest of stages involved in the considered running stage. The algorithms are performed for each collection of preprocessed CQI observations shown in Table 4.4. The list of parameters which are used to optimize each clustering algorithm performance is given in Table 4.5. By setting the number of maximum stages per run of about $N_{st/run}^{max} = 10$, the Lloyd algorithm is randomizing at each 10 stages the set of centers in order to fast-up the process of finding better solutions. The total number of stages for all heuristic algorithms is $N_{st}^{max} = 1000$. The minimum consecutive RDL must be small enough for the hybrid-SAST such as $RDL_C^{Min} = 0.1$

Table 4.5 The Parameter Settings of K-means Clustering Algorithms

Parameter Name	Parameter Description
Number of Max Stages per Runs ($N_{st/run}^{max}$)	10
Minimum CRDL and ARDL (RDL_C^{Min} , RDL_A^{Min})	0.1
Initial Acceptance Probability Better Solutions Pr_{Ac}	0.5
Initial Probability Transition Swap To Lloyd $Pr_{S\to L}$	0.5
Temperature Running Length (N_{Stages}^{Temp})	10
Temperature Reduction Factor (R_T)	0.95
Tunneling Parameter (γ_{ST})	0.02
Total Number of Stages (N_{st}^{max})	1000
Maximum Number of Swaps per Stage ($N_{SW/st}^{max}$)	1

due to the fact that only when $RDL_C\left[t_{st}\right] > 0.1$, then a new Lloyd stage is preferred to be performed instead of the Swap heuristic. Otherwise, a new center is swapped in/out without wasting one stage for the local search. The temperature running length is $N_{Stages}^{Temp} = 10$ due to the fact that when exceeding this threshold, the impact in the temperature value is reduced. During the temperature running length, the initial probabilities for non-better solution acceptance Pr_{Ac} or for transition from Swap to Lloyd $Pr_{S\to L}$ is set to 0.5 in order to provide equal chances of selection for both decisions. The tunneling factor is $\gamma_{ST} = 0.02$ which means that the global minimum is assured based on a time window of 50 stages. If the tunneling parameter is too large, then $Pr_{Ac} \approx 1$ and $Pr_{S\to L} \approx 1$, and if the tunneling parameter is too small, then the time window for the global solution detection is very reduced. Therefore, the setting of $\gamma_{ST} = 0.02$ represents a very good compromise. The cooling process needs to be achieved based on the number of maximum stages $N_{st}^{k} = 1000$ and then, the temperature reduction factor is $R_T = 0.95$. The number of maximum swaps for each stage is $N_{SW/st}^{max} = 1$ in order to detect the impact of each swapped center from the candidate list.

Figure 4.8 shows the performance of the proposed method for Top3 preprocessed CQI mass mode when the number of centers is $N_{CT} = 64$ and the system bandwidth is $\mathrm{BW} = 15\mathrm{MHz}$ .The hybrid SAST performs much better when

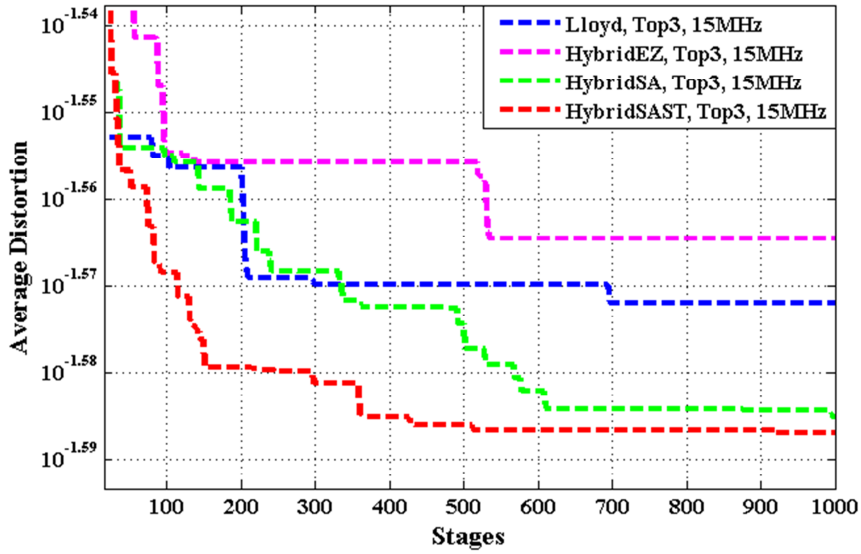


Fig. 4.8 The K-means Average Distortion for the Top 3 CQI Mass Mode with $N_{CT} = 64$

compared with the hybrid-SA from the best average distortion point of view. It can be seen that the hybrid-SAST achieves the minimum distortion after 500 stages, whereas the hybrid-SA needs more than 100 stages in order to achieve an appropriate performance. Other approaches such as Lloyd and hybrid-EZ are not suitable to find a better set of centers because of getting stuck in the local minima problem. However, hybrid-EZ outperforms the iterated Lloyd due to the fact that at the beginning of each running stage, one center is swapped from the set of candidate centers. Without having a precise control of Lloyd and Swap stages, the hybrid-EZ is not able to reach the minimum distortion level imposed by the hybrid-SAST. Similar behaviors are observed in Fig. 4.9 ($\mathrm{BW} = 10\mathrm{MHz}$) and Fig. 4.10 (BW=20MHz) for Top4 mass mode and Top5 mass mode, respectively, with the same number of preprocessed CQI clusters $\left(N_{CT} = 64\right)$. In both cases, the hybrid-SAST outperforms other considered methods.

In Figure 4.10, the minimum distortion is achieved by the hybrid-SAST after 550 stages whereas a comparable but not better performance is achieved by the hybrid-SA approach after 800 stages. It is clear that, by using the tunneling function, the selection decisions of Lloyd or Swap aim to find better sets of centers when compared with the hybrid-SA. The Swap's best average distortion is omitted from Figures 4.8, 4.9 and 4.10 due to its poor performance when compared with other existing clustering methods. The swap clustering can achieve

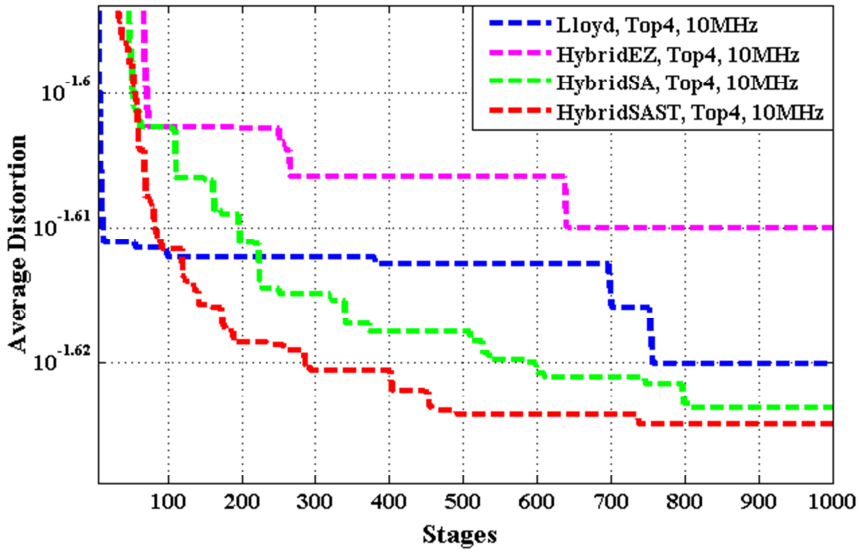


Fig. 4.9 The K-means Average Distortion for the Top4 CQI Mass Mode with $N_{CT} = 64$

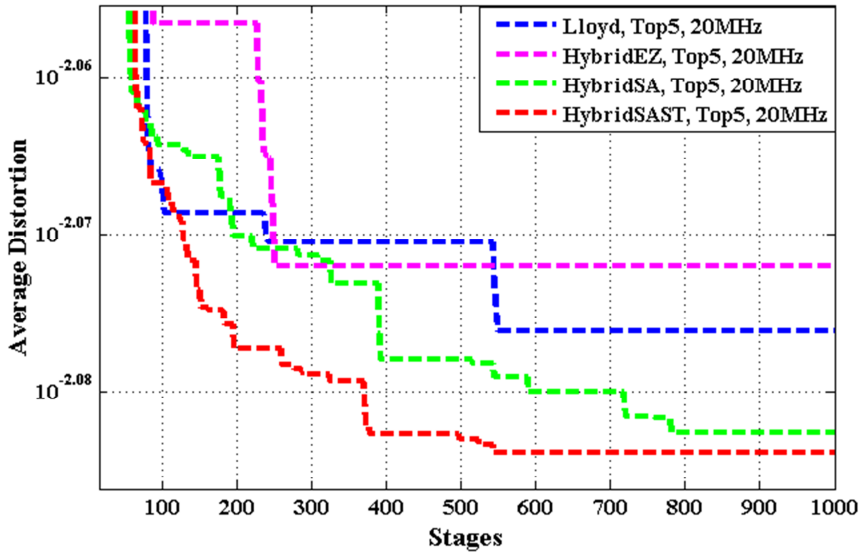


Fig. 4.10 The K-means Average Distortion for the Top 5 CQI Mass Mode with $N_{CT} = 64$

better performances after a larger number of stages when all the possible combinations of centers will be swapped in/out from the candidate list.

In order to highlight the importance of the proposed hybrid-SAST clustering algorithm, Appendix D presents the performance of the analyzed algorithms for each system bandwidth with different numbers of centers by using the top mass configuration models: Top3, Top4 and Top5. The obtained results show the best average distortion performance for stages 100, 250, 500, 750 and 1000 based on the percentage of the relative increase to the Lloyd algorithm. In

more than 70% of the simulation results, the hybrid-SAST performs better than other considered clustering algorithms. Due to the tunneling function computation, in most of the considered cases, the hybrid-SAST algorithm requires more computation time when compared with the hybrid-SA, but this fact is not a major issue since at this stage, only the offline learning step is performed.

4.7.3 Variable Number of K-Means Centers

In the previous sub-section, the best average distortion evolution over a variable number of stages and a constant number of centers were analyzed and compared. In order to find the proper number of k-means centers for the RBFNN generalization and classification, the number of centers variability must be considered. Through extensive simulation results, the impact of the number of centers N_{CT} is analyzed for each clustering algorithm, for each system bandwidth and for each considered top mass preprocessing scheme. The analyzed range for the numbers of centers is $N_{CT} = \{1,...,1024\}$, and the simulation parameters keep the same values as indicated in Table 4.5.

Figure 4.11.a shows the hybrid-SAST best average distortion for the Top3 mass mode preprocessing configuration which is achieved at the end of each 1000 stages for each $N_{CT} = 1,...,1024$ number of centers. Interestingly, the best average distortion is achieved in the case when the system bandwidth is BW=20MHz which has the largest CQI collection size. The best set of centers decreases in performance if the CQI data set size decreases. To conclude, by enlarging the population size of the CQI clusters, the probability of finding a better set of centers with lower distortions can be strongly improved. The Central Processing Unit (CPU) computation time for the same CQI data sets is presented in Fig. 4.11.b. As expected, the lowest collection size when the system bandwidth is BW=1.4MHz provides the best results when compared with other data sets for other bandwidths. However, the hybrid-SAST for BW=20MHz indicates a better performance than other bandwidths excepting the case when BW=1.4MHz. As mentioned earlier, the most time-consuming step in k-means clustering algorithms

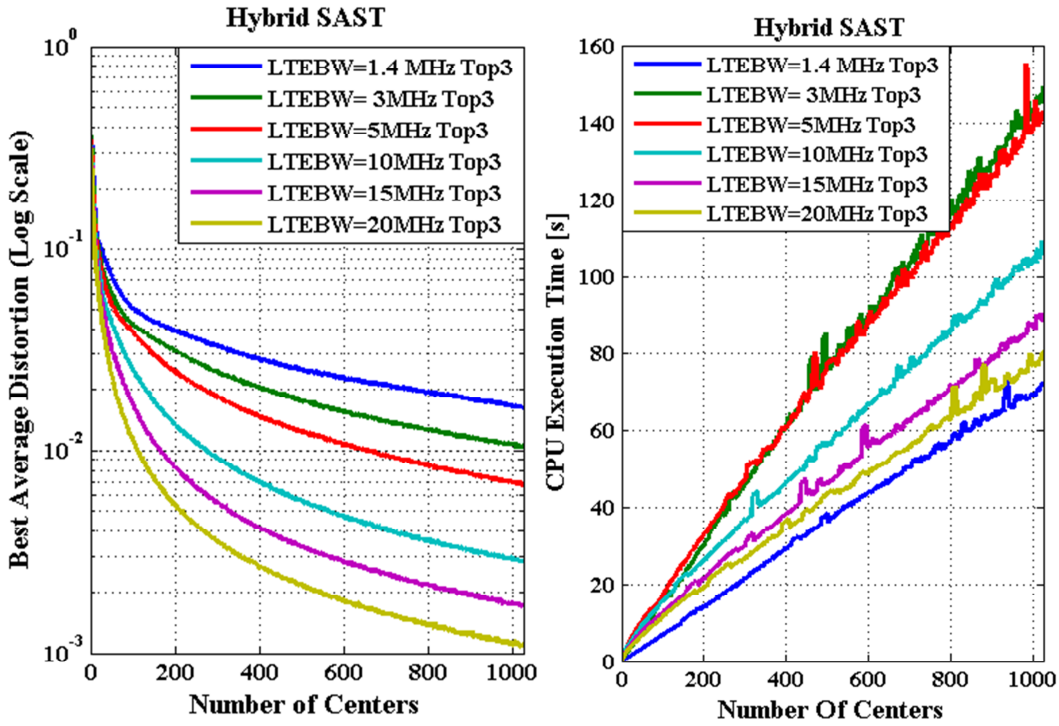


Fig. 4.11 a) The Best Average Distortion and b) The CPU Execution Time for Hybrid SAST under the Top 3 CQI Mass Mode

is the k-d tree computation. However, this aspect does not affect the overall comparison since the same step is performed for other data sets. Another reason is the neighborhood calculation for the Lloyd step. When the data set has proper geometrical characteristics, the neighbors can be reached very easily. It is the case of the data set when the system bandwidth is 20MHz. When the preprocessed CQI sets have un-proper geometrical properties, the neighborhood determination increases the system complexity. It is the case of the preprocessed CQI collections with 3MHz and 5MHz bandwidths.

The same computations are performed in Figure 4.12 for the Top4 CQI mass mode. The best average distortion for all preprocessed CQI sets is higher when compared with the Top3 case. The reasoning deals with two aspects: the collections contain 4 non-zero elements for each observation and the data set sizes are larger when compared with the Top3 case. In order to find better minimum distortions, the number of stages should be increased for each center point calculation. As expected, the CPU execution time becomes higher than in the previous case, but the result properties keep the same form.

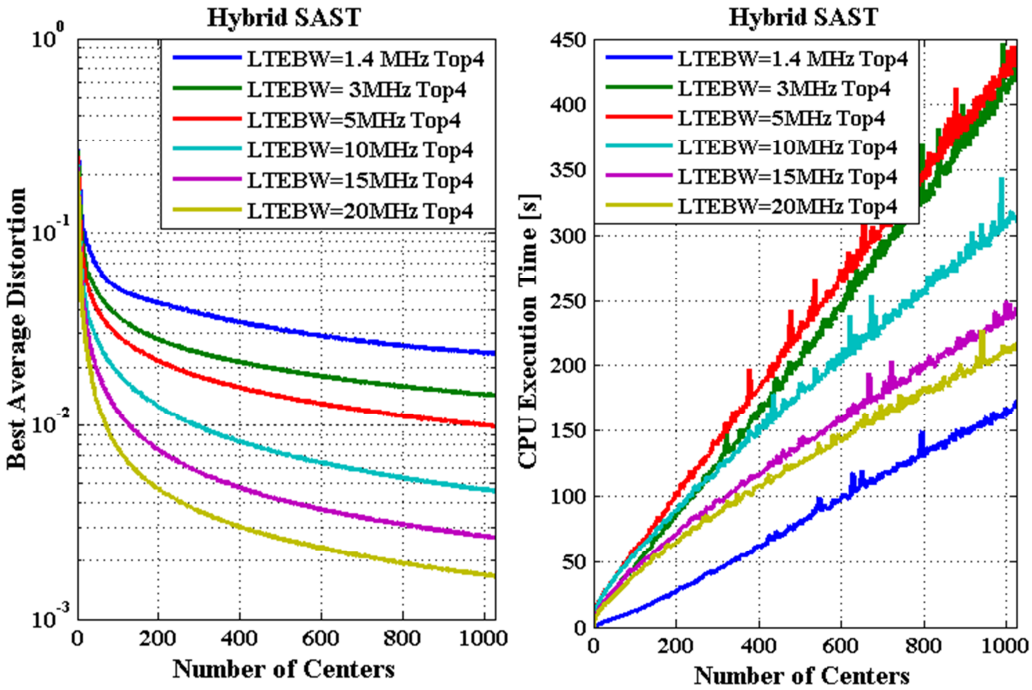


Fig. 4.12 a) The Best Average Distortion and b) The CPU Execution Time for Hybrid SAST under the Top 4 CQI Mass Mode

For the Top5 CQI mass mode (Fig. 4.13), the overall performance for the best average distortions and CPU computation time shows a larger degradation when compared with the Top3 and Top4 data sets. By increasing the number of non-zero elements of input observations, the hybrid-SAST method requires more stages to minimize the best average distortion and to find a better set of centers. This aspect implicitly affects the CPU execution time which becomes higher when the data set size increases. The above reasoning affects the RBFNN training and exploitation procedure since the computation complexity for each hidden node can be severely affected when the top CQI mass mode has a higher number of non-zero elements. In Chapter 6, the reassignment mass modes for Top3, Top4 and Top5 provide an adequate accuracy of the CQI state space classification in order to find the sustainable scheduling policies being focused on the NGMN requirement. The performances of other clustering algorithms for the considered CQI data sets are analyzed in Appendix E. It is very important to notice that the Swap algorithm provides the best performance from the CPU time point of view since this approach does not require the neighborhood solution calculation. The distortion is noisy due to the swapping procedure which is applied at each stage.

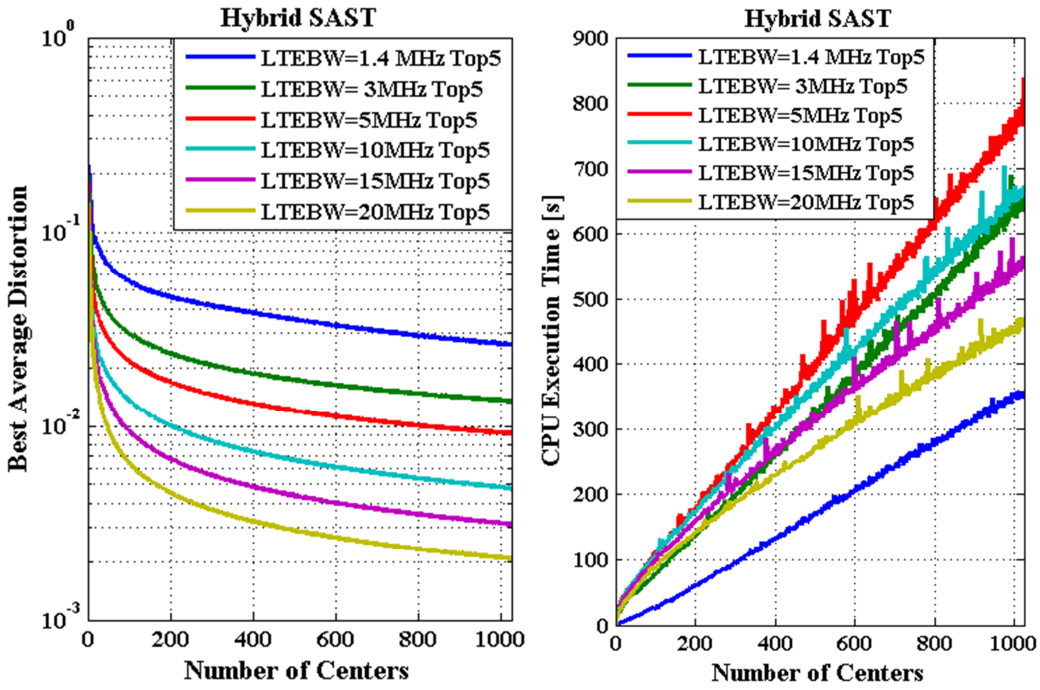


Fig. 4.13 a) The Best Average Distortion and b) The CPU Execution Time for Hybrid SAST under the Top 5 CQI Mass Mode

4.7.4 RBFNN Training/Validation Errors Based on Un-optimized Parameterization

The optimized sets of k-means centers obtained with the hybrid-SAST methodology are used by the RBFNN hidden layer. As mentioned earlier, the number of centers should be equivalent with the number of hidden nodes for the appropriate interpolation. The golden section of the proposed classification method aims to use two different sets of observations in the training stage such as validation and training sets. The number of observations from the validation sets for each configuration is shown in Table 4.4. The same SAST concept is used to decide the training or the validation observation as an input for the RBFNN structure. The parameter settings for the proposed methodology are presented in Table 4.6. The initial temperature is loaded based on the tunneling function for $N_{Epochs}^{Temp} = 50$ TTIs, and then, the temperature decreases gradually based on the simulated annealing scheduler. The initial acceptance probability for the training set is higher being set to Pr_{Tr} =0.8 in order to train the RBFNN weights more on

Table 4.6 RBFNN Parameter Settings with Un-optimized Learning Rate and Gaussian Weight

Parameter Name	Parameter Description
Validation Data Points Selection Method	SAST
Acceptance Probability for Training Set (Pr_{Tr})	0.8
Temperature Running Length (N_{Epochs}^{Temp})	50000
Temperature Reduction Factor (R_T)	0.99
Tunneling Parameter (γ_{ST})	0.1
Number of CQI Feedbacks at each TTI ($\lvert\mathcal{U}_t\rvert$)	1000
Total Number of Epochs (N_{Eph})	1000000 Epochs = 1000 TTIs
System Bandwidth (BW)	20 MHz
User Speed	120 Km/h
Fading Type	Jakes Model
PUCCH Model	Errorless with periodic CQI reporting mode
RBFNN Input Layer Function (φ_i)	Linear
RBFNN Hidden Layer Function	Gaussian
Number of Hidden Nodes (N_H)	N_{CT}
RBFNN Output Layer Function	Tangent Hyperbolic
Number of Output Nodes	$log_2(N_{CT})$
Learning Rate (η_{RBF})	0.1
Gaussian Weight (σ_{RBF})	10

the observations which are provided by the simulator instead of using the preprocessed observations from the validation set. In order to decrease the temperature uniformly based on the total number of epochs $N_{Eph}=10^6$, the temperature reduction factor is very high being set to $R_T=0.99$. Due to the higher number of epochs when compared with the SAST clustering algorithm, the tunneling parameter is $\gamma_{ST}=0.1$ in order to search the global minimum solution for the RBFNN errors by using a time window length of 10 epochs. The learning and Gaussian parameters are not optimized and the main focus in this section is to monitor the average RBFNN output error for different number of k-means centers for the system bandwidth of 20MHz. The RBFNN structure is coupled to the LTE CQI feedback module for the training set and a very large number of users $\lvert\mathcal{U}_t\rvert=1000$ with random initial positions and 120kmph speeds are used to fast-up the learning procedure. The Jakes multipath model is used in order to avoid the

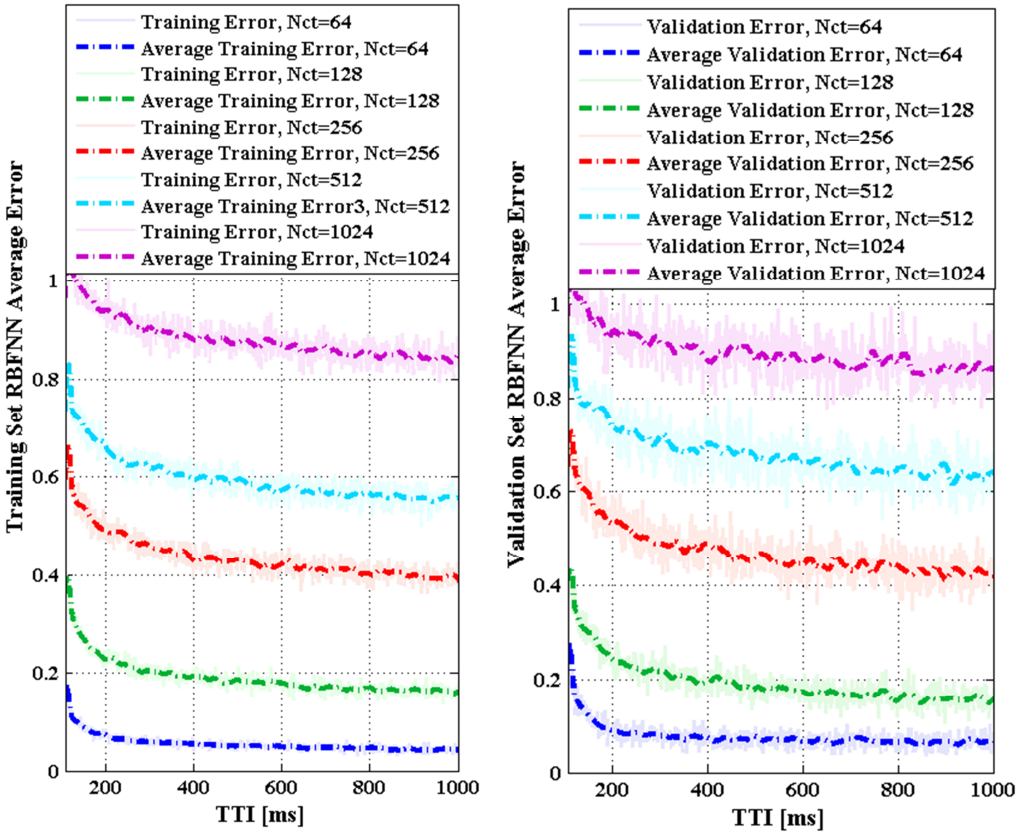


Fig. 4.14 a) RBFNN Training Errors and b) RBFNN Validation Errors for the Preprocessed Top 3 CQI Mass Mode

local minima problems and the CQI reporting mode is periodic at each TTI in order to maximize the number of observations for the RBFNN training set. It is important to note that the sets of RBFNN weights which minimize the average output error during the training stage are saved and loaded for the testing phase.

Figure 4.14 shows the impact of the considered learning rates (η_{RBF}) and Gaussian weights (σ_{RBF}) in the RBFNN output error for the Top3 CQI mass mode and for both types of training and validation sets. As expected, the training average error outperforms the validation set error due to the geometrical properties of the input data. Moreover, the error variation is higher for the validation sets due to the fact that each time when SAST decides to select a validation observation, the observation selection is purely random, leading to the non-uniformly geometrical properties. From Fig. 4.14, it can be concluded that the RBFNN output error increases if the number of hidden nodes increase. For a higher number of centers, the number of weights becomes considerably larger and

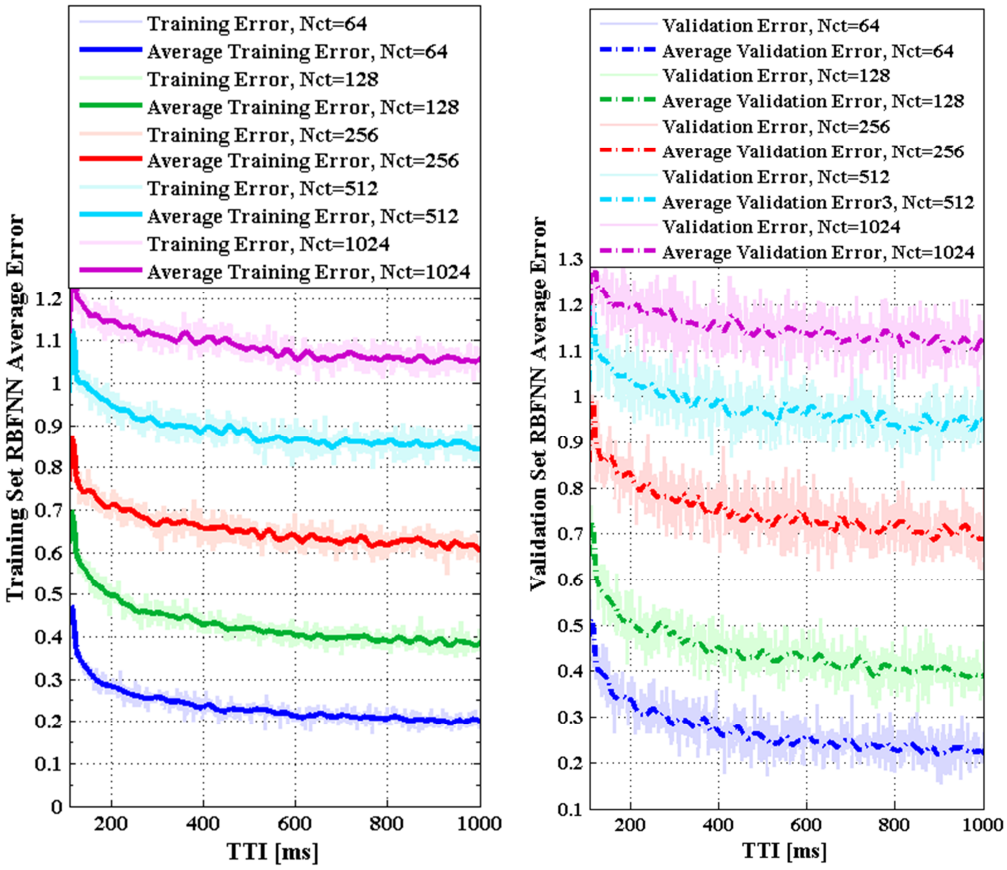


Fig. 4.15 a) RBFNN Training Errors and b) RBFNN Validation Errors for the Preprocessed Top 4 CQI Mass Mode

requires more input observations for the proper correction of the RBFNN weights. The same principles are exposed in Fig. 4.15 for the same system bandwidth with the Top4 CQI mass mode for a variable number of centers. The overall output error for both training and validation sets is degraded by about 0.15 when compared with the Top3 mass mode configuration. This fact is due to a larger number of non-zero elements in the preprocessed input state space which requires more time for a proper training. The error variation becomes higher especially for the validation set in which the randomization of the observation selection introduces noticeable noise in the output back-propagated error. These drawbacks are even more visible in Fig. 4.16 for the Top5 CQI mass mode when the preprocessed CQI collection size increases. It is easy to observe that the overall error performance is degraded by at least 0.35 in the error difference when compared with the Top3 mass mode. It is expectable that if the top configuration is increased, the RBFNN output error becomes larger. The maximum error limit

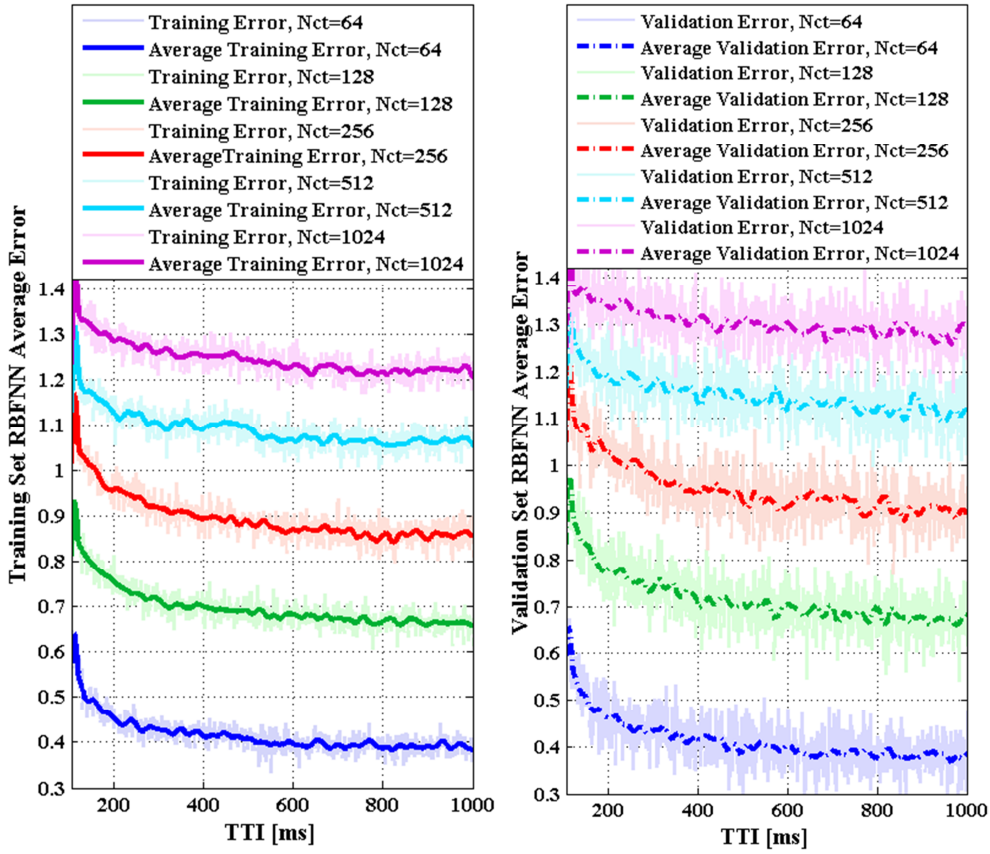


Fig. 4.16 a) RBFNN Training Errors and b) RBFNN Validation Errors for the Preprocessed Top 5 CQI Mass Mode

is 2, which means in fact that up than $\overline{E}\left(\mathcal{S}_{O}^{CQI,C},\mathcal{S}_{O,Patt}^{CQI,C}\right)=0.5$, the mean error level is not acceptable anymore for the CQI state space classification. The only way is to optimize the RBFNN functionality based on two parameters: *the learning rate and the Gaussian weights*. As shown earlier, when the number of centers increases, the average output error becomes larger. The main drawback of considering a larger number of centers implies the fact that the RBFNN structure should be trained for a longer time of simulation. The main advantage of a larger number of centers is the classification accuracy. With the higher classification accuracy, the regressed CQI state space has a better representation of the overall classified CQI state space. On the other hand, by increasing the number of hidden nodes, the computation complexity is much higher when compared with the traditional case of Top3 CQI mass mode with $N_{CT}=64$ number of centers. Therefore, the operator should decide which configuration fits better from the viewpoints of the system complexity and the accuracy performance.

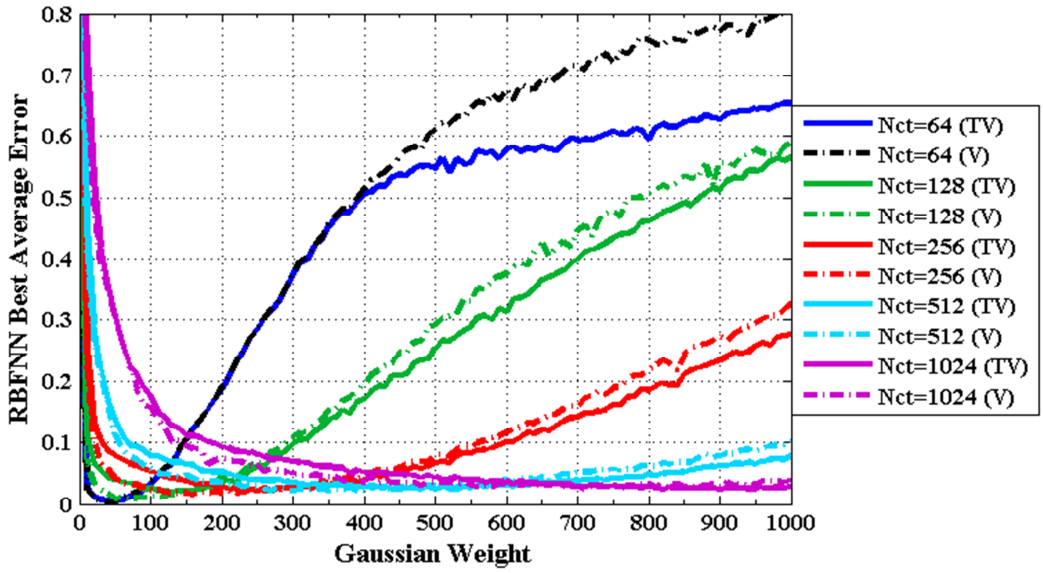


Fig. 4.17 The RBFNN Output Error over Variable σ_{RBF} and Constant $\eta_{RBF} = 0.1$ for the Preprocessed Top3 CQI Mass Mode

4.7.5 Optimization of RBFNN Parameters

Based on comprehensive simulation results, the optimal sets of learning and Gaussian parameters which minimize the RBFNN average output error are provided in this sub-section. The simulation results are performed for different settings of learning rates and Gaussian weights (η_{RBF} , σ_{RBF}) by using the same set of parameter values from Table 4.6 with a slight difference in which the number of SAST temperature running length becomes $N_{Epochs}^{Temp} = 200$. Figure 4.17 shows the impact of the Gaussian parameter on the output error computation for different numbers of hidden nodes. For all scenarios, the RBFNN average output error aims to decrease for some intervals and to increase beyond the considered intervals. In Fig. 4.17, for a number of $N_{CT} = 1024$ CQI centers, the evolution of the Gaussian parameter σ_{RBF} does not bring any improvement in the error performance after exceeding the threshold of $\sigma_{RBF} = 700$. However, by increasing the number of centers, the feasible interval which can guarantee the RBFNN average error minimization increases. In Figure 4.18 the same simulations are performed for the Top4 CQI mass mode. The RBFNN error is degraded when compared with the previous case, especially when the configuration of $N_{CT} = 1024$ is used due to the

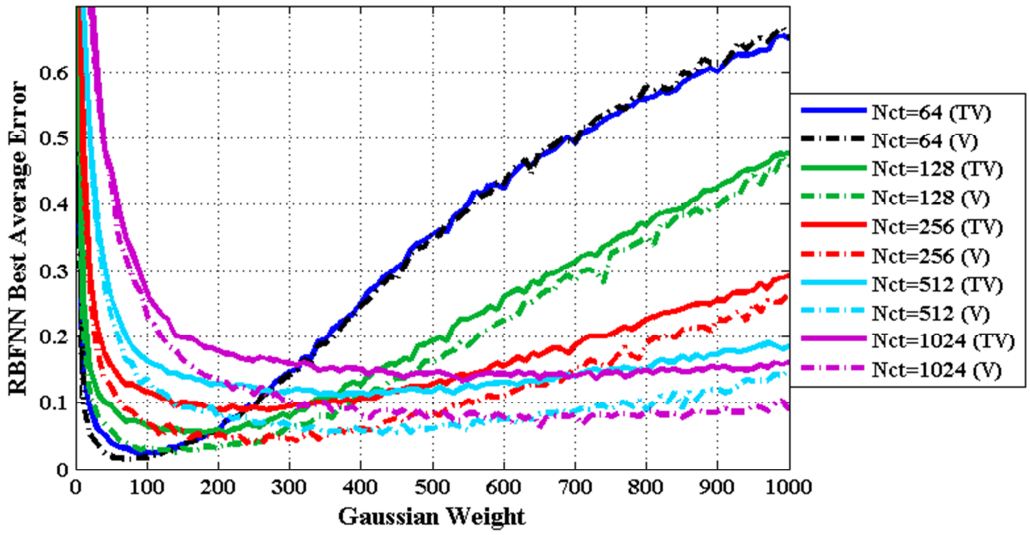


Fig. 4.18 The RBFNN Output Error over Variable σ_{RBF} and Constant $\eta_{RBF} = 0.1$ for the Preprocessed Top4 CQI Mass Mode

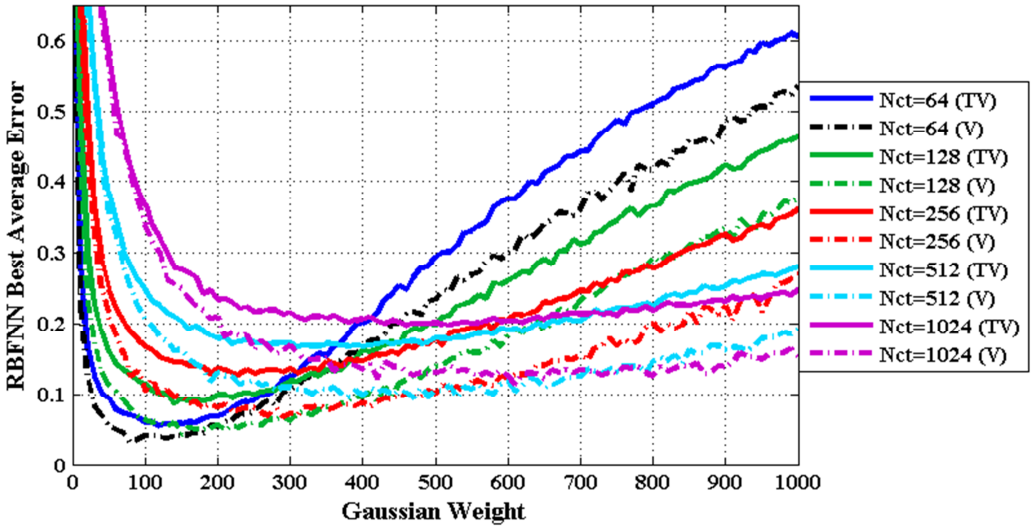


Fig. 4.19 The RBFNN Output Error over Variable σ_{RBF} and Constant $\eta_{RBF} = 0.1$ for the Preprocessed Top5 CQI Mass Mode

limitation on the maximum number of training epochs. The same behavior of average error is noticed in Fig. 4.19 for the Top5 CQI mass mode case. When the preprocessed top mass mode configuration increases, the feasible range of σ_{RBF} becomes larger for $N_{CT} \in \{64,128\}$ and more restrictive for $N_{CT} \in \{256,512,1024\}$.

In Figures 4.20, 4.21 and 4.22, the impact of variable learning rates on the RBFNN average error is analyzed for a constant Gaussian weight of $\sigma_{RBF} = 10$.

The variability of learning rates introduces opposite characteristics for the RBFNN configurations with $N_{CT} \in \{64,128\}$ number of centers when compared with the previous case of the Gaussian weight variability. In this sense, the feasibility range in which the configurations of $N_{CT} \in \{64,128,256,512,1024\}$ can find an optimal learning parameter η_{RBF} for a minimum RBFNN error becomes more restrictive when the number of non-zero elements in the preprocessed CQI

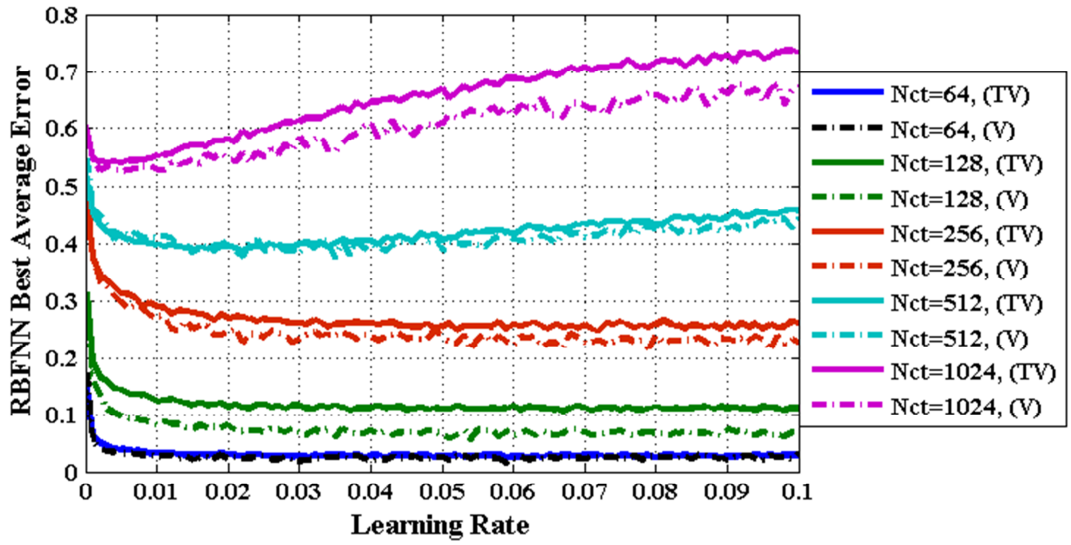


Fig. 4.20 The RBFNN Output Error over Variable η_{RBF} and Constant $\sigma_{RBF} = 10$ for the Preprocessed Top3 CQI Mass Mode

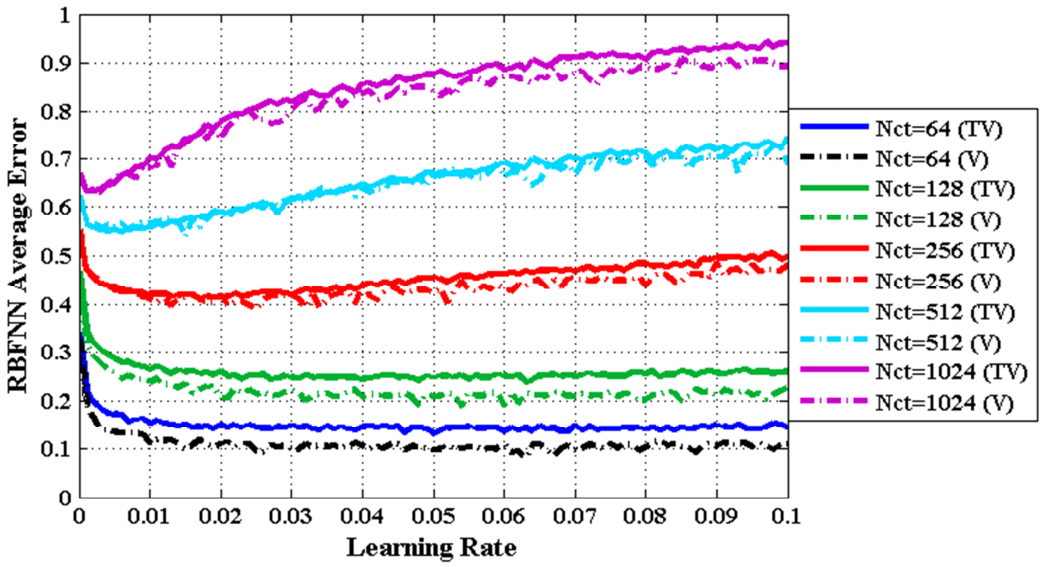


Fig. 4.21 The RBFNN Output Error over Variable η_{RBF} and Constant $\sigma_{RBF} = 10$ for the Preprocessed Top4 CQI Mass Mode

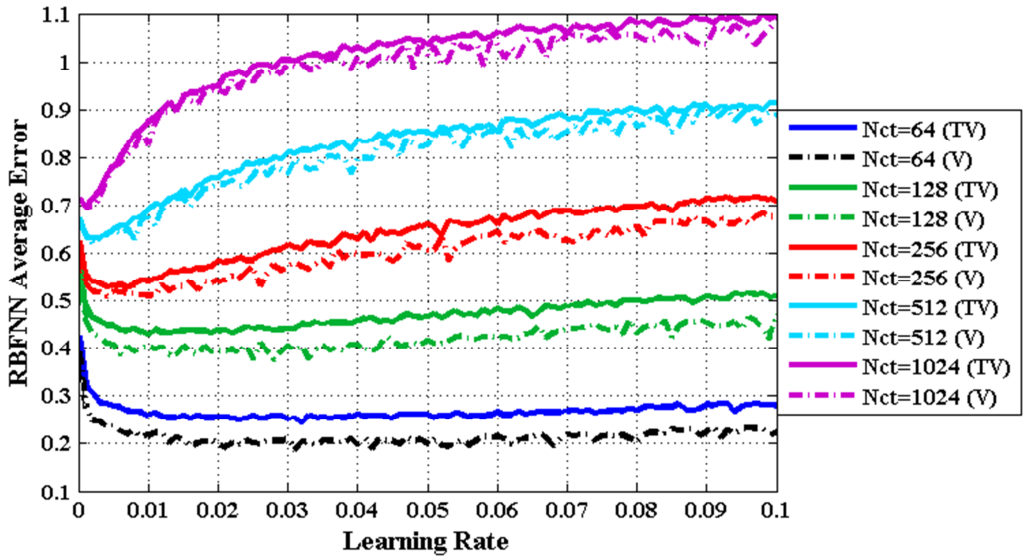


Fig. 4.22 The RBFNN Output Error over Variable η_{RBF} and Constant $\sigma_{RBF}=10$ for the Preprocessed Top5 CQI Mass Mode

observations increases. Based on the simulation results obtained in Figs. 4.20, 4.21 and 4.22, the learning parameter does not bring significant improvement in the RBFNN average error minimization. It is more convenient to adjust the Gaussian parameter based on a predefined learning parameter. Table 4.7 shows the optimized learning and Gaussian parameters which can be used in the CQI state space aggregation stage based on multiple classifier configurations. The parameters from Table 4.7 are used in the CQI state space aggregation for the DSR-SMOO/CMOO problems to be presented in Chapters 6 and 7.

4.7.6 RBFNN Training/Validation Errors Based on the Optimized Parameterization

The simulation results from Sub-section 4.7.4 are reloaded by using the optimal set of learning and Gaussian parameters presented in Table 4.7. As seen from Fig. 4.23, the RBFNN average output training error for the Top3 mass mode and for $N_{CT}=1024$ number of centers decreases with approximately 0.8 after 1000 TTIs when compared with the simulation results from Figure 4.14.a. The average errors for both training and validation sets show an improved performance when compared with the un-optimized case from Figure 4.14.

Table 4.7 An Optimal Set of Learning and Gaussian Parameters for BW=20MHz

Top / N_{CT}	Top3		Top4		Top5	
	η_{RBF}	σ_{RBF}	η_{RBF}	σ_{RBF}	η_{RBF}	σ_{RBF}
64	0.089	50	0.05	90	0.032	120
128	0.075	130	0.063	180	0.01	180
256	0.072	270	0.021	210	0.007	230
512	0.022	440	0.006	370	0.002	310
1024	0.005	1000	0.001	540	0.001	460

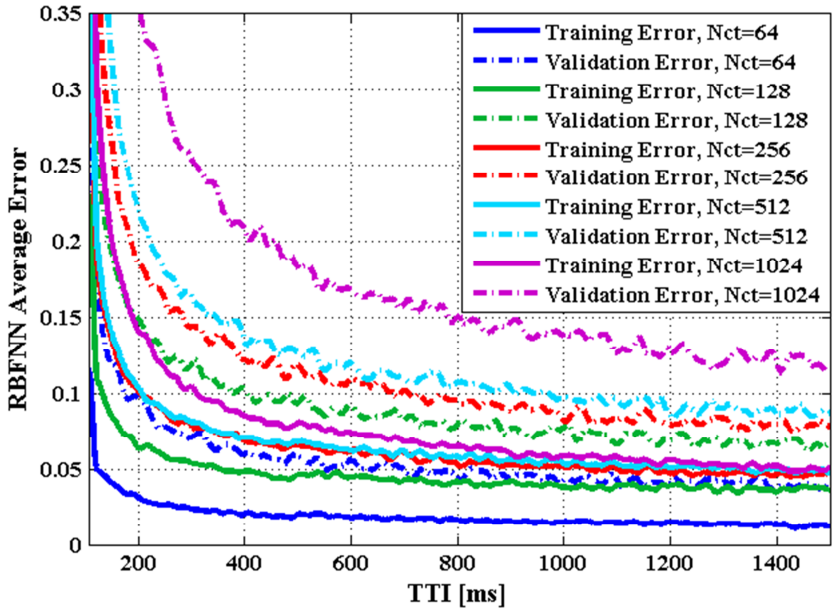


Fig. 4.23 RBFNN Training/Validation Average Errors with Optimal Parameterization for the Preprocessed Top 3 CQI Mass Mode

In Figure 4.23, the classification procedure for the entire set of considered configurations is achieved with the minimum RBFNN average error since $\overline{E}\left(\mathcal{S}_{O}^{CQI,C},\mathcal{S}_{O,Patt}^{CQI,C}\right)<0.1$. It is observable that in this case, the validation average error is much higher when compared with the training average error. This concept is not applied for other configurations such as Top4 and Top5 mass modes as indicated in Figures 4.24 and 4.25. The explanation behind this phenomenon is directly regarded to the collection of the preprocessed CQI observations. When the number of observations from the collected preprocessed CQI set is much lower than the entire preprocessed CQI state space size, the RBFNN validation error can decrease with a higher rate when compared with the training errors as indicated by the Top4 and Top5 configurations. This issue does not represent a

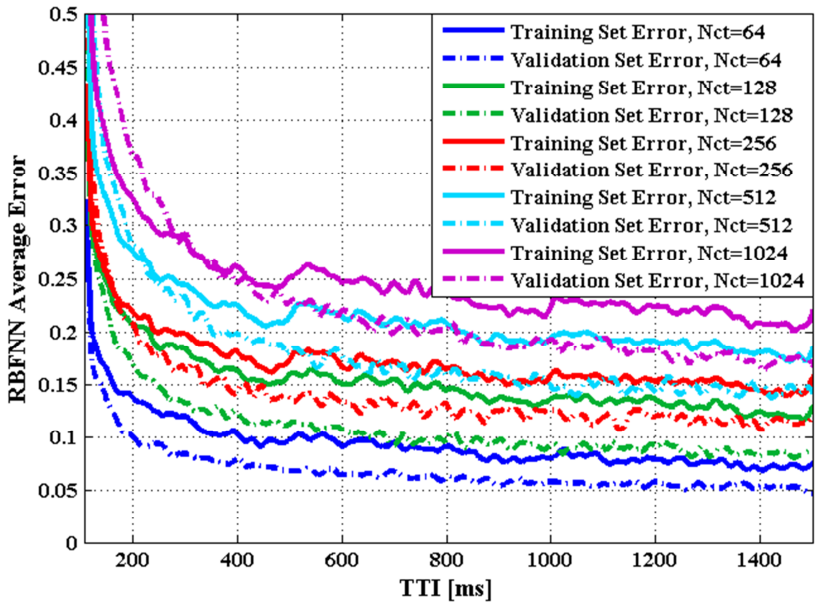


Fig. 4.24 RBFNN Training/Validation Average Errors with Optimal Parameterization for the Preprocessed Top 4 CQI Mass Mode

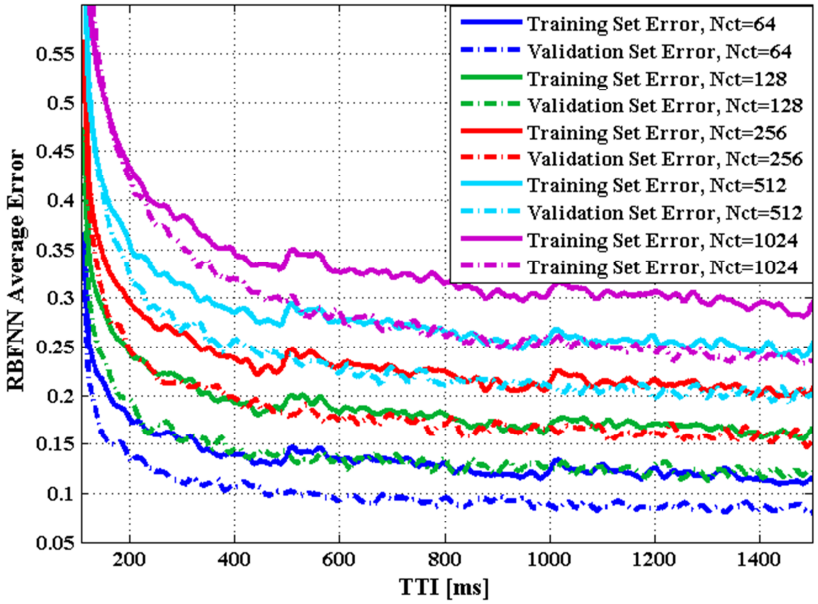


Fig. 4.25 RBFNN Training/Validation Average Errors with Optimal Parameterization for the Preprocessed Top 5 CQI Mass Mode

major problem from the RBFNN weight correction point of view since the sets of weights are saved based on the total average error which comprises the training and validation errors. The minimum training average error for the Top5 mass mode and $N_{CT} = 1024$ number of centers is less than 0.35 which represents an acceptable limit to set up the saved set of weights for the RBFNN testing stage.

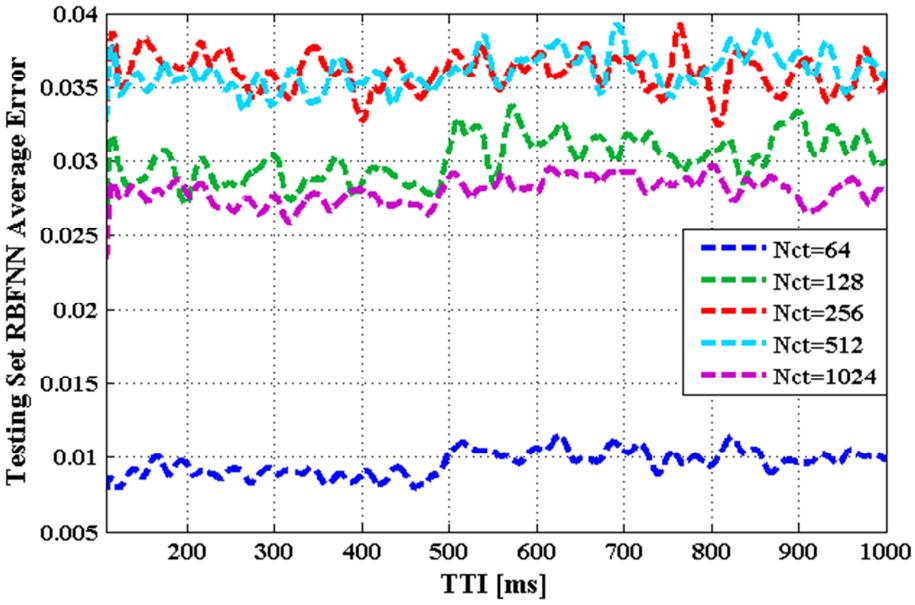


Fig. 4.26 RBFNN Testing Average Errors with Optimal Parameterization for the Preprocessed Top 3 CQI Mass Mode

4.7.7 RBFNN Testing Errors Based on the Optimized Parameterization

The RBFNN testing stage exploits the optimal set of weights obtained in the previous section. At this level, only the feed-forward procedure is performed in which the preprocessed CQI observations for different top mass mode configurations are classified based on the corrected set of weights obtained through the back-propagation procedure from the training stage. The testing stage considers different initial user positions from the training stage in order to highlight the veracity of the trained set of weights. The rest of the simulation parameters keep similar values to those in the previous scenarios. The minimum RBFNN average output error is achieved for the particular case when the number of centers is set to $N_{CT} = 64$. As expected, when the number of hidden nodes increases, the RBFNN classification structure requires more time epochs to minimize the average error of its output layer. For the considered number of training epochs, the performance of the average output error decreases gradually when the number of hidden nodes increases. But even under these circumstances, for a large number of k-means centers, the RBFNN structure fits very well for the

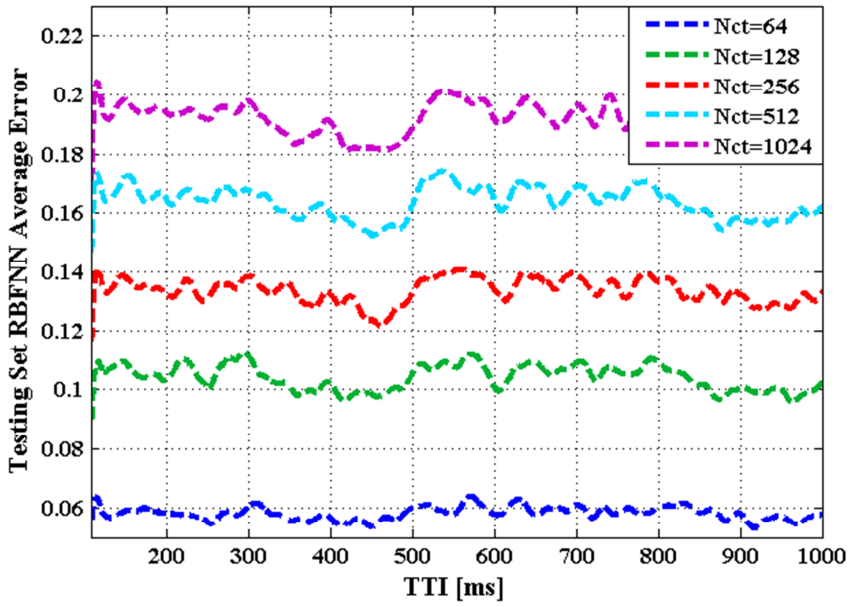


Fig. 4.27 RBFNN Testing Average Errors with Optimal Parameterization for the Preprocessed Top 4 CQI Mass Mode

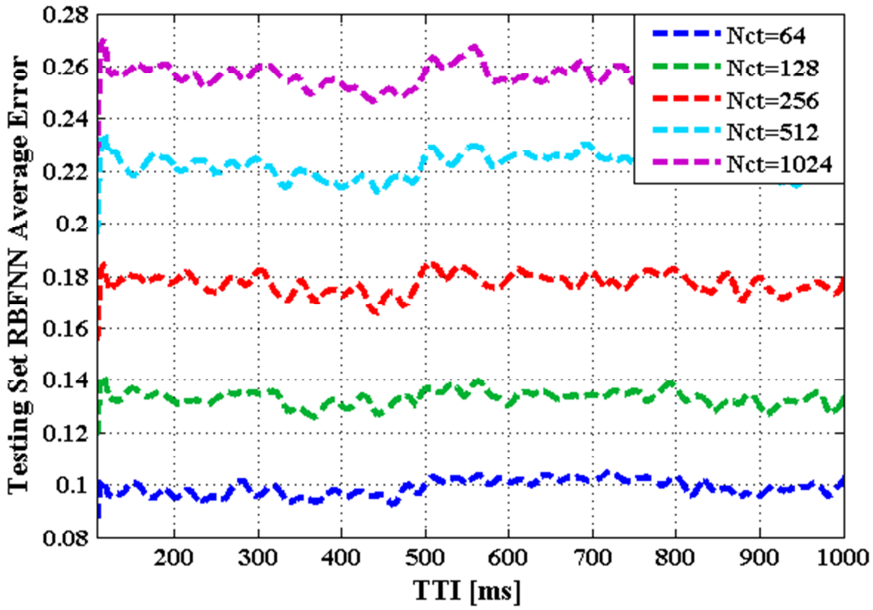


Fig. 4.28 RBFNN Testing Average Errors with Optimal Parameterization for the Preprocessed Top 5 CQI Mass Mode

CQI classification purpose, achieving an output error of about less than 0.3 for all considered cases as shown in Figures 4.26, 4.27 and 4.28. Based on the trained RBFNN structure, the classified outputs can be further exploited as a component of the controller input state space after performing the regression stage. The sets of sustainable scheduling policies obtained in Chapter 6 and Chapter 7 are trained

based on the regressed CQI state space being obtained at each TTI by using the principles proposed in this chapter.

4.8 Summary

The original scheduler state space aggregation is considered to be crucial for the DSR-SMOO/CMOO scheduling problems due to its high dimensionality. The controllable state space can be transformed in a more compactable version by using statistical models among different active bearers. The CQI state space is the most important element of the uncontrollable scheduler state space. An innovative methodology of CQI state space aggregation is proposed in this chapter. The preprocessing stage is applied in order to bring the initial CQI state space to a more compacted version which does not depend directly on the system bandwidth. The classification stage is performed in order to classify the preprocessed CQI observations in different channel quality classes. The set of optimal preprocessed CQI centers is obtained based on the proposed hybrid-SAST meta-heuristic method which is able to outperform the other existing clustering methods by minimizing the best average distortion between the collected CQI data points and the obtained k-means centers. The RBFNN structure is proposed for an accurate classification of the preprocessed CQI observations which is not included by the k-means data point collection. The proposed RBFNN architecture aims to select validation or training observations based on the same SAST schedule which monitors the consecutive epoch errors. The feed-forward and backward propagation modules are used by the RBFNN structure to minimize the average output error. Based on multiple configurations of preprocessing schemes and number of k-means centers with optimal parameterization, the RBFNN avoids the local minima and the over-fitting problems. The simulation results indicate that the SAST based RBFNN with feed-forward and backward propagation is very suitable in the CQI state space classification by minimizing the average RBFNN output error. The regression stage is performed based on the obtained classified state space that is able to provide statistical information about the channel qualities for the input state space of the LTE/LTE-A scheduler controller.

Chapter 5

LTE Packet Scheduling Based on Reinforcement Learning

5.1 Chapter Outline

The LTE scheduling procedure can be modeled by using the MDP processes in which the current scheduling decision depends only on the previous one. The temporal difference learning is used to select scheduling rules TTI-by-TTI by rewarding the previous applied rules (actions). RL approaches reinforce the target values TTI-by-TTI in order to maximize over the time the accumulated rewards for different visited controller states. The contribution of this chapter is to apply the existing RL approaches to the DSR-SMOO/CMOO problems. Due to the fact that the aggregate controller state space is continuous, the MLPNN function approximations are used in order to generalize the state-action values and the state values for the unvisited states. The MLPNN weights are trained by using the gradient descent principle. This chapter proposes a novel LTE MAC scheduler architecture based on intelligent controller which can be used under different modes. When the homogeneous traffic type is scheduled, the controller makes use of two agents with specific cooperation for the fairness and QoS objectives. The fairness agent learns how to meet the fairness feasibility state whereas the QoS controller is responsible for reaching the controller state feasibility from the viewpoint of the entire set of scheduling objectives.

5.2 The LTE-A Scheduler Controller and the RRM Environment Interface

The LTE scheduler controller and the RRM environment interface aim to manage the interaction between the decision-making (which is called by the intelligent controller) and the RRM entities such as the packet scheduler and the multi-objective evaluator. The interaction procedure between the controller and the environment comprises three main stages:

1. At each TTI *t,* a new scheduler state $\mathcal{S}_t^S$ is sensed and aggregate to a more compactable form for the scheduler controller such as $\mathcal{S}_t^C$ based on the principles presented in Chapter 4.
2. By using transition probabilities, a new action $\mathcal{A}_t^a\left(\mathcal{S}_t^C\right) \in \mathcal{A}$ is selected, where $a = 1,..,\left|\mathcal{A}\right|$ and $\mathcal{A}$ represents the action set that can be finite or infinite $\mathcal{A} \in \mathbb{R}^{D[\mathcal{A}]}$.
3. When performing the action $\mathcal{A}_t^a\left(\mathcal{S}_t^C\right)$, the environment evolves to the next aggregate state $\mathcal{S}_{t+1}^C$ in the next TTI. As a result of its action $\mathcal{A}_t^a\left(\mathcal{S}_t^C\right)$ performed in the previous TTI, the controller receives from the RRM environment the reward value $\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) \in \mathbb{R}_{[-1,1]}$. As mentioned in Chapter 3, the reward function indicates at TTI *t*+1 the performance of action $\mathcal{A}_t^a$ being applied in state $\mathcal{S}_t^C$ at TTI *t*.

The details of the considered elements involved in the interaction process between the controller and other RRM entities are introduced in the following sub-sections.

5.2.1 The LTE Scheduler Controller State Space

The LTE controller state space aggregation is analyzed in Chapter 4 and implies the reduction of its space without losing any relevant information which can be useful for the controller's decision $\mathcal{A}_t$. Different aggregation functions are applied in different regions of the original scheduler state space in order to

compress the controllable and uncontrollable subspaces. However, even under these circumstances, the obtained controller state space $\mathcal{S}_t^C$ is *continuous* and *infinite*. Therefore, it is impossible for the controller to sweep all the possible states, and another processing unit is necessary in order to approximate the unvisited states based on the experience obtained when visited by other ones.

If the RRM environment is known or partially known, the transition probabilities between the controller states can be defined [203]. The transition function represents the probability of reaching state $\mathcal{S}_{t+1}^C = \mathcal{S}' \in \mathcal{S}^C$ when action $\mathcal{A}_t^a = \mathcal{A}^a \in \mathcal{A}$, $\forall a = 1,..,N_A$ is applied in the previous state $\mathcal{S}_t^C = \mathcal{S} \in \mathcal{S}^C$. The transition probability is defined by Eq. 5.1.

$$\mathcal{P}_{\mathcal{S},\mathcal{A}^a}^{\mathcal{S}'} = p\left(\mathcal{S}'|\mathcal{S},\mathcal{A}^a\right) = Pr\left\{\mathcal{S}_{t+1}^C = \mathcal{S}'|\mathcal{S}_t^C = \mathcal{S}, \mathcal{A}_t^a = \mathcal{A}^a\right\} \tag{5.1}$$

where $p\left(\mathcal{S}'|\mathcal{S},\mathcal{A}^a\right)$ is the probability of getting the state $\mathcal{S}'$ when the action $\mathcal{A}^a$ is applied in current state $\mathcal{S}$. Unfortunately, the RRM environment in LTE scheduling cannot offer a model based on the transition probabilities. Therefore, $\mathcal{P}_{\mathcal{S},\mathcal{A}^a}^{\mathcal{S}'}$ is omitted from the RL algorithms in order to reflect the fact that the RRM environment is totally unknown for the LTE scheduler controller.

5.2.2 The LTE Controller Action Space

The controller action $\mathcal{A}_t^a$ can take discrete or real (continuous) values. At the same time, the controller action can be a vector with discrete or real values. The controller action is mapped by the MUTI entity into a proper marginal utility function or scheduling rule. Then, the MU decision vector takes the form of:

$$\mathcal{A}_t^a \xrightarrow{MUTI} c[t] = \left\{c_{o,w_o}[t], o = 1,..,\left|\mathcal{O}\right|, w_o = 1,..,\left|\mathcal{PU}_o\right|\right\}, \forall a = 1,...,N_A \tag{5.2}$$

where $N_A = \left|\mathcal{A}\right|$ is the total number of discrete actions if and only if $\mathcal{A}$ is finite. When the action is a real vector, just a part of the output decision vector is used for the marginal utility selection. Other parameters from the action vector can be

used to fine tune some parameters from the selected MU weight such as fairness parameters or to inform the RAC module for the acceptance of new bearers.

5.2.3 The Reward Function

The reward function represents a quality measure of applying a certain action $\mathcal{A}_t^a$ for a given controller state $\mathcal{S}_t^C$ following the notation $\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right)$. Obviously, the reward should be a function which depends on the performance of the DSR-SMOO/CMOO problem satisfaction in the short term purpose (averaged over a relative short TTI window) in order to sense the immediate effect of applying the considered controller action. The reward function should be connected somehow with the optimality condition of the scheduler controller state. The reward function should be designed in such a way that a maximum reward value represents the optimal region where the scheduler should operate as long as possible. Then, the general reward function can be defined as follows:

$$\mathcal{RW}: \mathcal{S}^C \times \mathcal{A} \to \mathbb{R} \quad \mathcal{RW}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t\right) = \mathcal{RW}^o_{\mathcal{S}_t^C, w_o}\left[t+1\right] = \sum_{o^*=1}^{|\mathcal{O}|} \delta_{o^*} \cdot \mathcal{RW}_{t+1}^{o^*} \qquad (5.3)$$

where $\mathcal{RW}^o_{w_o}\left[t+1\right]$ is the aggregate reward value at TTI t+1 when the decision variable $c_{o,w_o}\left[t\right]$ has been applied in the previous TTI t $\forall o \in \mathcal{O}$, $\forall w_o \in \mathcal{PU}_o$, and $\mathcal{RW}_{t+1}^{o^*}$ is the reward value for the objective $o^* \in \mathcal{O}$.

5.2.4 Controller Policies

At each TTI, the controller maps the input state in probabilities of selecting specific actions. The mapping procedure between states and actions is called the controller or the agent *policy*. The policy is denoted by $\pi_{\mathcal{A}}\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right)$ and represents the probability of selecting action $\mathcal{A}_t^a$ when the input state is $\mathcal{S}_t^C$. Therefore, the definition domain for the controller policy is $\pi_{\mathcal{A}}: \mathcal{S}^C \times \mathcal{A} \to \left[0,1\right]$. Let us define the controller action selected at TTI t $\mathcal{A}_t^a$ and the rest of non-

selected actions are such as $\mathcal{A}_t^{na}\left(\mathcal{S}_t^C\right) \in \mathcal{A}$, $\forall na = 1,..,|\mathcal{A}|$, $na \neq a$. The controller policy $\pi_{\mathcal{A}}$ is considered to be *stochastic* in state $\mathcal{S}_t^C$ if and only if the policy value for action $\mathcal{A}_t^a$ is $\pi_{\mathcal{A}}\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right) = 1$ and for the non-selected action $\mathcal{A}_t^{na}$, the value is $\pi_{\mathcal{A}}\left(\mathcal{A}_t^{na} \mid \mathcal{S}_t^C\right) = 0$, $\forall na = 1,..,|\mathcal{A}|$, $na \neq a$ [203]. If the policy $\pi_{\mathcal{A},t} = \pi_{\mathcal{A}}$ is not changing over the time, then the controller policy is *stationary*.

The marginal utility policy $\pi_{dU}\left(c_{o,w_o}[t] \mid \mathcal{S}_t^S\right)$ represents the binary version of the controller policy $\pi_{\mathcal{A}}\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right)$ which in fact selects a scheduling decision variable $c_{o,w_o}[t]$ based on the controller action $\mathcal{A}_t^a$. That is, the controller policy provides an action and MUTI selects a proper scheduling rule based on the mapping procedure between the controller and the scheduler actions. A stationary controller policy implies the SSR-SMOO/CMOO problems when one single scheduling rule is applied across the whole scheduling period, whereas the stochastic one involves the novel DSR-SMOO/CMOO combinatorial problems.

5.3 LTE Scheduling as a Markov Decision Process

The discrete-time Markov Decision Process (MDP) [204], [205], [206] is used together with the RL algorithms in order to manage the interaction between controller and the RRM environment. At the basic principle, MDP serves as input for a given RL algorithm and aims to find the optimal policy [207]. Then, the role of the LTE controller together with the RL procedure is to solve a given MDP problem. The MDP problem aims to extract an action $\mathcal{A}_t^a$ from a given policy $\pi_{\mathcal{A}}$ which is learned so far. Being concentrated on finding the optimality of the DSR-SMOO/CMOO problems and under the fact that the number of TTIs (steps) tends to infinite, the MDP problem is considered to have an *infinite horizon*.

With the infinite horizon, the state value should be bounded in order to avoid some convergence problems [206]. This is the reason why the DSR-SMOO/CMOO MDP problems should be *discounted*. In real practice, a proper discount factor is very hard to be found, and for this reason, additional processing

units for the state values are proposed for the DSR-SMOO/CMOO MDP problems in order to ensure the policy convergence. Then, the MDP is defined based on the tuple $\left(\mathcal{S}^C, \mathcal{A}, \mathcal{RW}, \mathcal{P}^{\mathcal{S}'}_{\mathcal{S},\mathcal{A}^a}, \gamma\right)$ where γ is the discount factor introduced in Sub-section 3.6.1.3 from Chapter 3. Based on the DSR-SMOO/ CMOO problem type, other parameters can be introduced in the specific MDP.

Due to the LTE scheduling problem nature, the MDP problem considers some important characteristics which are listed below:

- The state space is *infinite* and *multi-dimensional*. This fact implies the impossibility of sweeping all the possible states, and thus, the requirement of interworking with a function approximation in order to obtain the state-action values becomes mandatory. The controller state space is considered *fully observable*.
- In order to solve the DSR-SMOO problems focusing on the fairness objective, the controller action space is considered to be *continuous* and *multidimensional*.
- The DSR-SMOO/CMOO MDP is *stochastic* in the sense that the rewards are noisy for different objectives, which in fact implies the stochastic nature of the reward function.
- The DSR-SMOO/CMOO MDP problem is *not always episodic*, which means that in some conditions, the optimal state cannot be accessed directly from any controller state. Each state can be accessed from any other states by using a finite number of TTIs. For this reason, the DSR-SMOO/CMOO MDP problem is considered to be *ergodic*.
- A terminal state for the DSR-SMOO/CMOO MDP problems is similar to the feasible or optimal scheduler state when the user rates are optimally allocated by satisfying at the same time the QoS requirements (e.g., GBR, fairness, PLR(PDR), HoL packet delay and stability). Moreover, the termination condition of the DSR-SMOO/CMOO problems is considered to be random. In some conditions, the terminal state cannot be reached and depends on the number of active users (to be detailed in Chapter 7). This is another reason why the DSR-SMOO/CMOO MDP problem is stochastic.

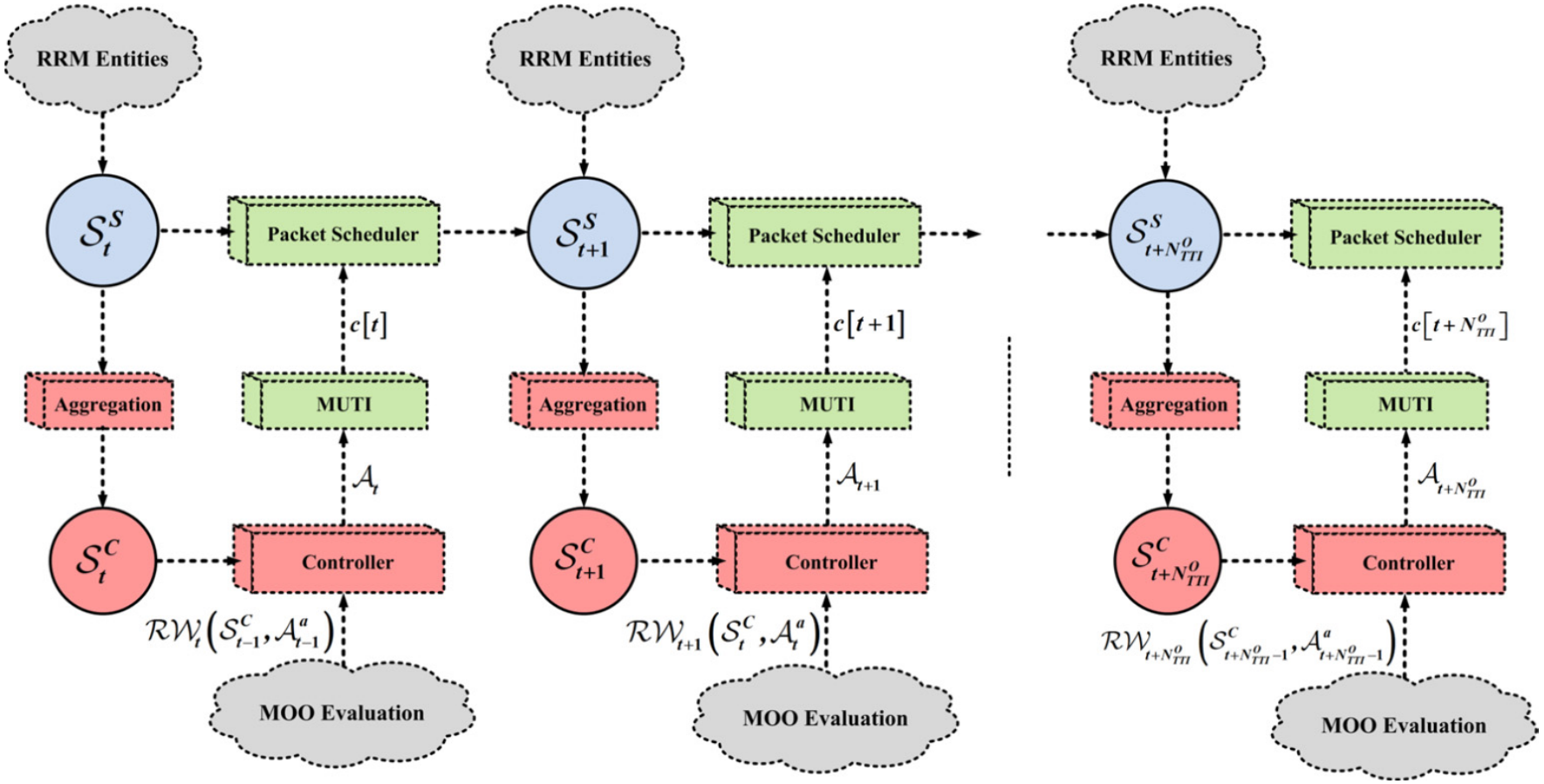


Fig. 5.1 The DSR-SMOO/CMOO MDP Principle

All of the characteristics listed above are taken into account in order to design the LTE controller for the decision on the scheduling rules. Figure 5.1 illustrates the interaction between the environment and the scheduler controller under the MDP representation. Due to the fact that the RRM environment (including the MOO evaluation) is completely unknown for the LTE scheduler controller, the state transition probabilities $\mathcal{P}_{\mathcal{S}_t^C,\mathcal{A}_t^a}^{\mathcal{S}_{t+1}^C}$ are omitted from this representation. At TTI t+1, the MOO evaluation senses how far or close the current state $\mathcal{S}_{t+1}^C$ is from the optimal one $\mathcal{S}_{t+N_{TTI}^O}^C$, where N_{TTI}^O is the number of TTIs when the system is declared optimal by considering TTI t as the initial time instant. Then, a reward value $\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)$ is computed. After the aggregation procedure, the controller applies another action based on the learned policy.

The MDP problem implies by definition to respect the Markov property [203], [208]: *A stochastic process has a Markov property if and only if the distribution of the current state depends only on the previous state*. This formulation is equivalent with the affirmation that the reward function and the transition probabilities depend only on the previous state and do not depend on all the states that the controller/scheduler visited in the past. Then, the DSR-SMOO/CMOO MDP model can be defined as $\left(\mathcal{S}_t^C,\mathcal{A}_t^a,\mathcal{RW}_{t+1},\mathcal{S}_{t+1}^C,\mathcal{A}_{t+1}^a,\gamma\right)$, where the unknown variable is the action to be applied at TTI t+1 $\mathcal{A}_{t+1}^a$.

In the reinforcement learning, an agent is a policy $\pi_{a,t}$ of selecting the action $\mathcal{A}_t^a$ from the action set $\mathcal{A}$. *For the DSR-CMOO with the homogeneous traffic type, an agent represents a policy of selecting a given scheduling rule*. In the case of the heterogeneous traffic type, the controller is used to coordinate a group of agents such as Multi-Agent RL (MARL). Each agent has its own policy which can be performed in the scheduling domain only and only if that policy is selected at different TTIs. For simplicity, the DSR-SMOO/CMOO MDP model is analyzed with the premise that the controller coordinates more than one agent. More sophisticated architectures are introduced in the upcoming sections for different QoS targets and then, in Chapter 8, for different traffic types.

5.4 The Coordinated Multi-Agent RL Based LTE Scheduling Policies

In general, the MARL systems can bring many benefits due to the distributed nature of the solution [209], [210]. Sharing the experience in multi-agent systems can help and speed-up the learning procedure for each agent which is involved in the exploration process [210]. When one agent fails in the multi-agent system, the responsibility can be taken by other agents [210]. But the MARL approach is not able to eliminate the main problems which are met in the LTE scheduling such as the scheduler space dimensionality and the exploration/ exploitation tradeoff which are considered the main limitations of the proposed approach. The MARL approach brings new problems in LTE scheduling, such as the learning-goal of each agent, the convergence to the Nash equilibrium and the need for coordination and cooperation [209], [210], [211]. The Nash equilibrium is the most used stability requirement when the multi-agent systems are used [212]. In this sense, many approaches are concentrated on the convergence to the MARL Nash coordinated equilibrium [213-217].

Alongside the exposed characteristics, the MARL approach is preferred in LTE scheduling because the current approach requires a high degree of scalability. The fairness policy differs from other QoS objectives policies in the sense that the reward, the state spaces and the action spaces are totally different. On the other hand, the heterogeneous traffic classes can be scheduled by simply selecting a class or multiple classes to be served. The MARL algorithms can be classified based on the cooperation and coordination methodologies such as:

1. **Fully Coordination**: The agents receive the same reward, and a central controller follows the MDP form exposed in Sub-section 5.3. The role of the centralized controlling is to coordinate the agent actions in order to maximize the long term reward. Then, the types of coordination can be *free-cooperation* which assumes that a joint action is considered to be optimal for each learning epoch [218], and for each agent, *specific cooperation* in which the global RL algorithm is divided into specific local

RL algorithms of different agents with reduced dimensions [219] and *indirect cooperation* which considers the joint action selection. Each agent learns the model about other agents [216].

2. **Explicit Coordination**: The state space and the reward function may be different for each agent, and the action selection is based on negotiation techniques [220].
3. **Competitive Agents**: When one agent action is selected to be performed, other agents act in the opposite way by minimizing its benefit [218].
4. **Mixed Cooperative/Competitive Agents**: By considering the particular case of fully cooperative tasks, agents may encounter the situations where they can be in conflict of interests [221].

The DSR-SMOO/CMOO learning goal is to maximize the MOO function under each situation. In this sense, each agent should bring its own contribution to the newest state without punishing other agent actions. The most suitable architecture for the DSR-SMOO/CMOO approaches for both homogeneous and heterogeneous traffic types is the MARL based on full coordination. For the homogenous traffic type, there is a specific cooperation between fairness agent and QoS agent since the QoS agents decide when the fairness agent should perform. For the heterogeneous traffic type, a distributed architecture is needed in order to coordinate a group of QoS agents specific for each traffic type. As mentioned earlier, a centralized controller is required in this case in order to select the best agent action at each TTI. The Distributed RL (DRL) approach is used in the DSR-SMOO/CMOO scheduling as a fully coordinated MARL without cooperation. The DRL methodology is used with great success in many deterministic and stochastic problems [222], [223], [224], [225]. The master agent selects the action based on its own policy, and based on its action, the corresponding slave agent is chosen. The role of the slave agent is to apply actions which will be converted in different scheduling disciplines by the MUTI entity. Basically, the slave agent uses the combination of fairness and QoS agents. To conclude, when the DRL approach is used for the heterogeneous traffic type, the full coordination architecture is used for the selection of the slave agents and the specific cooperation is performed for the fairness and QoS agents belonging to the selected

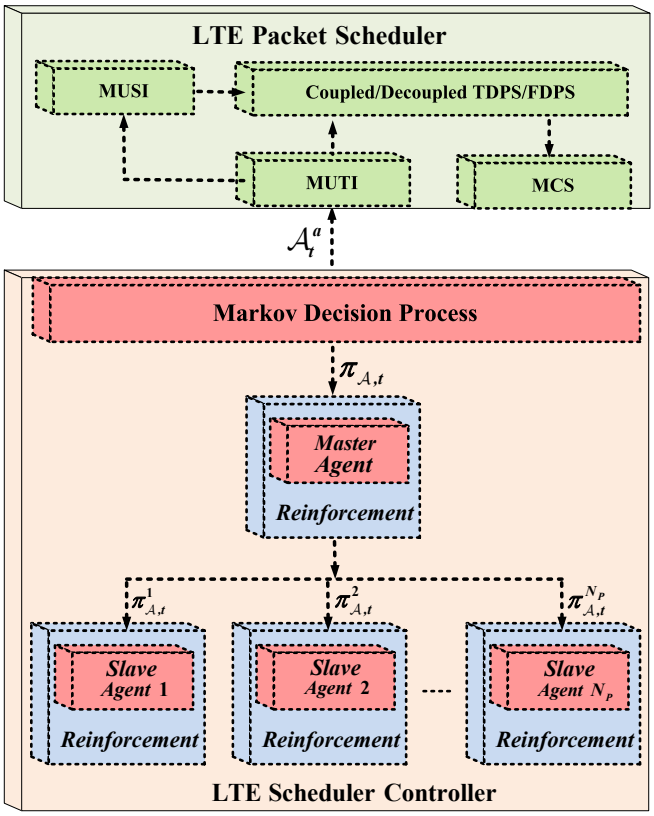


Fig. 5.2 The Distributed RL in Coordinated/Cooperation-Free Multi-Agent Systems

slave agent. As shown in Fig. 5.2, the LTE controller is designed to manage a group of agents who uses the same MDP characteristics. Therefore, this chapter aims to analyze three novel architectures as listed below:

1. **DSR-SMOO based on a Single RL Agent**: Different RL algorithms are presented and analyzed in order to be used in different QoS objectives.
2. **DSR-CMOO based on two RL Agents**: For the fairness objective, the single agent with continuous action is preferred. Two types of agents are implemented in this sense such as fairness and QoS agents.
3. **DSR-CMOO with the heterogeneous traffic type based on the Hierarchical MARL Approach**: A master agent is required to face the overall controller state space, and based on its action, the sub-space for a given priority class is transmitted to the fairness and QoS slave agents. The master agent is granted based on the total reward, and the slave agents are granted based on their particular performance subject to fairness and QoS constraints (to be detailed in Chapter 8).

5.5 Reinforcement Learning Principles in LTE Scheduling

Based on the MDP problem which is received at each TTI, the LTE scheduler controller has to *explore* and to *exploit* the aggregate scheduler state space in order to select a specific agent action that maximizes the aggregate reward. The *exploration stage* is responsible for selecting better actions in the future in order to improve the controller policy $\pi_{\mathcal{A},t}$. On the opposite pole, in the *exploitation stage*, the controller uses the known policy in order to extract a particular action for a given state that has the greatest amount of approximated accumulated rewards. The way how to combine the exploration and exploitation stages in order to improve or to use the existing policies captures a big interest in the proposed approaches and differs from one RL algorithm to another. Based on the exploration policy, it may be possible that the LTE scheduler controller spends more time on a given part of the controller state $\mathcal{S}_t^C$. For these reasons, the agents can lose the optimality of their actions in other regions of the controller state, and some RL algorithms will require an additional stage between exploration and exploitation known as the *experience replay* (ER) stage.

The RL algorithms work with state-action and state values in order to evaluate and to improve their policies. As mentioned earlier, the state space is continuous, and then, the state-action and state values can be obtained by using non-linear function approximations such as MLPNN functions. The clustering methods are not suitable for the controller state space classification since the precision of the state elements is crucial in determining the state-action values.

5.5.1 State and State-Action Values

As discussed in Sub-section 3.6.1.3, under a given MDP problem and a given controller policy $\pi_{\mathcal{A},t}$, the state value at TTI *t* can be estimated as follows:

$$V^{\pi_{\mathcal{A}}}\left(\mathcal{S}_t^C\right) = \mathbb{E}_{\pi_{\mathcal{A}}}\left\{\sum_{nt=1}^{\infty}\gamma^{nt-1}\cdot\mathcal{RW}_{t+nt}\left(\mathcal{S}_{t+nt-1}^C\right)\right\} \tag{5.4}$$

where γ represents the discount factor and $\mathbb{E}_{\pi_{\mathcal{A}}}\{\cdot\}$ is the estimation operator under a given controller policy $\pi_{\mathcal{A},t}$. Basically, the state value $V^{\pi_{\mathcal{A}}}\left(\mathcal{S}_t^C\right)$ follows the trajectory imposed by $\pi_{\mathcal{A},t}$ based on the accumulated rewards in the future states. Then, the controller aims to maximize this value for any policy such that:

$$V^*\left(\mathcal{S}_t^C\right) = \max_{\pi_a} V^{\pi_{\mathcal{A}}}\left(\mathcal{S}_t^C\right) \tag{5.5}$$

where the main goal is to find the optimal policy $\pi_{\mathcal{A}}^*$ in order to solve the DSR-SMOO/CMOO problem such as $V^{\pi_{\mathcal{A}}^*} = V^*$. Taking into account the recursion expressed in Eq. 3.64, the well-known Bellman equation is obtained [226]:

$$V^{\pi_{\mathcal{A}}}\left(\mathcal{S}_t^C\right) = \sum_{a=1}^{|\mathcal{A}|} \pi_{\mathcal{A}}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) \sum_{\mathcal{S}' \in \mathcal{S}_{t+1}^C} P_{\mathcal{S}_t^C, \mathcal{A}_t^a}^{\mathcal{S}'} \left(\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) + \gamma \cdot V^{\pi_{\mathcal{A}}}\left(\mathcal{S}'\right)\right) \tag{5.6}$$

It is assumed in Eq. 5.6 that the state and action sets have finite sizes. Then, the optimal state value is:

$$V^*\left(\mathcal{S}_t^C\right) = \max_a \left\{ \sum_{\mathcal{S}' \in \mathcal{S}_{t+1}^C} \mathcal{P}_{\mathcal{S}_{t+1}^C, \mathcal{A}_t^a}^{\mathcal{S}'} \left[\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) + \gamma \cdot V^*\left(\mathcal{S}'\right)\right] \right\} \tag{5.7}$$

Based on Eq. 5.7, the optimal controller state value at TTI *t* can be updated by selecting the best agent action which can provide the highest instantaneous reward in the scheduling time instant *t*+1.

Equivalent with the controller state value, the action value for a given state can be calculated by using the Bellman recursive representation [226]. For historical reasons, the state-action value at TTI *t*+1 is expressed by $Q_{t+1}^{\pi_{\mathcal{A}}}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right)$ and the updating equation is illustrated bellow [226]:

$$Q_{t+1}^{\pi_{\mathcal{A}}}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) = \sum_{\mathcal{S}' \in \mathcal{S}_{t+1}^C} \mathcal{P}_{\mathcal{S}_t^C, \mathcal{A}_t^a}^{\mathcal{S}'} \left\{ \mathcal{RW}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) + \gamma \cdot \sum_{a'=1}^{|\mathcal{A}|} \pi_{\mathcal{A}}\left(\mathcal{S}', \mathcal{A}_{t+1}^{a'}\right) \cdot Q_t^{\pi_{\mathcal{A}}}\left(\mathcal{S}', \mathcal{A}_{t+1}^{a'}\right) \right\} \tag{5.8}$$

The difference of Eq. 5.8 when compared with the state value function is the fact that the policy is located inside of the expectations of the future rewards. The

optimal action–state value $Q_{t+1}^{*}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)$ can be defined as follows:

$$Q_{t+1}^{*}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)=\sum_{\mathcal{S}'\in\mathcal{S}_{t+1}^C}\mathcal{P}_{\mathcal{S}_t^C,\mathcal{A}_t^a}^{\mathcal{S}'}\left\{\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)+\gamma\cdot\underset{a'}{max}\,Q_t^{*}\left(\mathcal{S}',\mathcal{A}_{t+1}^{a'}\right)\right\} \quad (5.9)$$

In Equations 5.7 and 5.9, it is assumed that the RRM environment model is perfectly known by having the states transition probabilities $\mathcal{P}_{\mathcal{S}_t^C,\mathcal{A}_t^a}^{\mathcal{S}'}$ and the state and action spaces are considered to be finite. By applying the mathematical models such as contraction mapping, the state and state-action values can be iterated [203]. Then, the dynamic programing can be applied since the model is known and the values can be updated [42]. However, the principle of dynamic programming cannot be applied in the DSR-SMOO/CMOO MDP problems since the RRM model is totally unknown.

Assuming that the RRM environment is unknown and the exploration of DSR-SMOO/CMOO MDP problems is episodic, the Monte Carlo method can be used. The state-action values represent the expected accumulated reward (sum of discounted rewards) starting from any initial state (Eq. 5.4). The MT principle starts with the premise that in the first update of the state-action value, the sum of discounted rewards from the initial state until the end of episode is used. Let us define t_{ep} as the start of the new episode, T_{ep} as the end of the episode where $t_{ep}<t<T_{ep}$. The state value can be updated at each episode in order to obtain a much better estimation of the state value based on the followed policy. Then, for the MT purpose, Equation 5.6 becomes [42]:

$$V_{t_{ep+1}}\left(\mathcal{S}_t^C\right)=V_{t_{ep}}\left(\mathcal{S}_t^C\right)+\eta_t^V\left(\mathcal{S}_t^C\right)\cdot\left[\sum_{nt=1}^{T_{ep}-1}\gamma^{nt-1}\cdot\mathcal{RW}_{t+nt}\left(\mathcal{S}_{t+nt-1}^C\right)-V_{t_{ep}}\left(\mathcal{S}_t^C\right)\right] \quad (5.10)$$

where $\eta_t^V\left(\mathcal{S}_t^C\right)$ is the learning rate for the value function which is used in order to mitigate the noise effect which may appear on the transition between states [203]. The learning parameter can be set based on the number of visits of each state assuming that the state space is still discrete or it can be approximated by using many trials based on the simulation results when the state space is continuous. The state-action values can be updated in a similar way, as indicated in Eq. 5.11:

$$Q_{t_{ep+1}}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)=Q_{t_{ep}}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)+\alpha_t^Q\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)\cdot \cdot\left\{\sum_{nt=1}^{T_{ep}-1}\left[\gamma^{nt-1}\cdot\mathcal{RW}_{t+nt}\left(\mathcal{S}_{t+nt-1}^C,\mathcal{A}_{t+nt-1}^a\right)\right]-Q_{t_{ep}}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)\right\} \tag{5.11}$$

As mentioned earlier, the DSR-SMOO/CMOO MDP problems are not episodic in all situations, making the MT approach unsuitable for the current purpose. Alongside this problem, the variance of the accumulated reward of the MT method is considerable [203]. In LTE scheduling, the main focus is to achieve as many updates as possible whereas the episodic updates reduce consistently the number of updates and implicitly the learning speed for a given number of TTIs.

5.5.2 Temporal Difference Learning

The temporal difference learning plays a central role in reinforcement learning. It represents a combination between Monte Carlo methods and Dynamic Programming (D-P) principles [42]. The D-P approach considers the problem of a perfect knowledge of the environment and contains a set of algorithms that is in charge of finding the optimal values, actions and policies. In contrast, in the case of Monte Carlo methods, the environment is unknown and the updates are achieved at the beginning of each episode. To conclude, *in the TD approach, the RRM environment is unknown and the state-action values are updated TTI-by-TTI*. By providing the immediate reward TTI-by-TTI, the variance of the accumulated reward is considerable lower [42], [203]. By imposing $T_{ep}\to\infty$ and $t=t_{ep}$, the state and state-action updates from Eq. 5.10 and 5.11 become:

$$V_{t+1}\left(\mathcal{S}_t^C\right)=V_t\left(\mathcal{S}_t^C\right)+\eta_t^V\left(\mathcal{S}_t^C\right)\cdot\left[\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C\right)+\gamma\cdot V_t\left(\mathcal{S}_{t+1}^C\right)-V_t\left(\mathcal{S}_t^C\right)\right] \tag{5.12}$$

$$Q_{t+1}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)=Q_t\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)+\eta_t^Q\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)\cdot \cdot\left[\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)+\gamma\cdot\max_{a'}Q_t\left(\mathcal{S}_{t+1}^C,\mathcal{A}_{t+1}^{a'}\right)-Q_t\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)\right] \tag{5.13}$$

where,

$$E_{t+1}^V\left(\mathcal{S}_t^C\right)=\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C\right)+\gamma\cdot V_t\left(\mathcal{S}_{t+1}^C\right)-V_t\left(\mathcal{S}_t^C\right) \tag{5.14}$$

$$E_{t+1}^{Q}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)=\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)+\gamma\cdot\max_{a'}Q_t\left(\mathcal{S}_{t+1}^C,\mathcal{A}_{t+1}^{a'}\right)-Q_t\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right) \quad (5.15)$$

are the TD errors of state and state-action values, respectively. Then, the main role of Equations 5.12 and 5.13 is to minimize the errors E_{t+1}^{V} and E_{t+1}^{Q} such that $\lim_{t\to\infty}V_t=V^{\pi_A}$ and $\lim_{t\to\infty}Q_t=Q^{\pi_A}$ with the learning rate properties: $\sum_{t=0}^{t\to\infty}\eta_t^{VQ}=\infty$ and $\sum_{t=0}^{t\to\infty}\left(\eta_t^{VQ}\right)^2<\infty$ according to [227]. The TD updating rule from Eq. 5.13 is known as the Q-learning algorithm [228]. The considered RL algorithm from Eq. 5.13 follows the tabular form since the controller state $\mathcal{S}_t^C$ is assumed to be discrete and thus, it can be stored in predefined tables. For the DSR-SMOO/CMOO MDP problems, this approach becomes unsuitable. Therefore, a function approximation should be considered in order to predict the state value $V_{t+1}\left(\mathcal{S}_t^C\right)$ and the state-action value $Q_{t+1}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)$ at TTI t+1 based on a multidimensional and continuous controller state space.

5.5.3 The Approximate RL in LTE Scheduling

The RL tabular representation for the DSR-SMOO/CMOO MDP problems can be achieved by using a preprocessing stage in which the continuous controller state space $\mathcal{S}_t^C$ can be clustered in a similar way to the pre-processed CQI classification. As a result, the precision of the input state space is reduced considerably, which is unacceptable for the DSR-SMOO/CMOO targets that are considered to be sensitive to the input state spaces. Therefore, the controller input state should keep the integrity of its features, making, in this way, the tabular RL unsuitable for the DSR-SMOO/CMOO MDP problems. In this sense, the state space approximation methods should be used in order to construct a predictive model for the state value $V_{t+1}\left(\mathcal{S}_t^C\right)$ and for the state-action value $Q_{t+1}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)$ for unseen continuous state-action pairs. In a general case of RL, for a given set of input data, there are some target outputs. The main problem refers to the non-stationary behavior of the target value due to the fact that the LTE scheduling is a stochastic process and the policy is varying during the exploration stage. In other

words, the approximation function should be flexible enough to follow the direction of the policy. There are several methods which can be used in the LTE scheduling for updating such approximation functions such as gradient descent, particle swarm optimization, simulated annealing or evolutionary algorithms. For the current purpose, the *gradient descent algorithm* principle, which was already discussed in Chapter 4 for the CQI state space classification, is used to fine tune the parameters of the scheduling rule function approximations.

In general, there are two types of approximations which can be used in RL, namely, such as linear function and non-linear function approximations. The linear approximations consist of a linear combination of the features in the input state space. Since the linear approximations depend on the features of the input state, the features and the quality of the input data are very important. In LTE scheduling, finding the features in the input state space implies an additional stage in which the data should be prepared and the features should be extracted. This issue is intolerable due to the time constraint imposed by the scheduling procedure. Even if the linear function has some convergence properties in the RL algorithms as shown in [229], [230], the non-linear function approximations are preferred to be used for the state and state-action predictions in the DSR-SMOO/CMOO MDP problems. The non-linear function approximations can be modeled as a MLPNN structure as proposed in [231].

The principle of the RBFNN generalization with feed-forward and back-propagation based on the gradient descent principle is explained in details in Sub-section 4.5.2 from Chapter 4. The same principle is used in the MLPNN functions for the state and state-action approximation but under a more generalized form.

The MLPNN structure is defined by the weight matrix $\mathcal{W}_l$ and by the activation functions φ_l, $l=1,..,N_L^{MLP}$ where N_L^{MLP} is the number of MLPNN layers (including the input and output layers). As seen in Sub-section 4.5.2 from Chapter 4, the sets of activation functions are known and the weight matrix should be updated and corrected through the gradient descent method. The weight matrix $\mathcal{W}_l$ dimension between two layers $(l,l+1)$ takes the value of $\left(N_N^l+1\right)\times N_N^{l+1}$,

where N_N^l is the number of nodes per layer l and the value of l (one) indicates the bias point. Given different numbers of nodes for each layer, the original controller state space $\mathcal{S}_t^C$ takes different dimensions according to the number of nodes N_N^l. Let us define the controller state space at the output nodes of layer l plus bias point such as $\mathcal{S}_{t,l}^{C,O}$. Then, the output state space of layer $l+1$ becomes:

$$\mathcal{S}_{t,l+1}^{C,O} = \varphi_{l+1}\left(\mathcal{W}_l^T \cdot \mathcal{S}_{t,l}^{C,O}\right) \tag{5.16}$$

where $\varphi_{l+1} = \left[\varphi_{l+1,1},...,\varphi_{l+1,N_N^l}\right]$. Then, the MLPNN function under the generalized form can be defined as expressed by Eq. 5.17:

$$\mathcal{F}_{MLP}\left(\mathcal{W}_t, \mathcal{S}_t^C\right) = \varphi_{N_L^{MLP}}\left(\varphi_{N_L^{MLP}-1}\left(...\varphi_l\left(...\varphi_1\left(\mathcal{W}_1, \mathcal{S}_t^C\right)...\right)...\right)\right) \tag{5.17}$$

where $\mathcal{W}_t = \left[\mathcal{W}_{t,1},...,\mathcal{W}_{t,N_L^{MLP}}\right]$ represents $N_L^{MLP}-2$ number of weights matrices. When $N_L^{MLP}=3$ implies the number of hidden layers $N_H^{MLP}=1$ and the number of weights which has to be adjusted at each TTI is $\left(N_N^3+N_N^1+1\right)\times N_N^2+N_N^3$. The activation functions used for the MLPNN function approximations are discussed based on each considered scenario in Chapters 6 and 7.

The MLPNN non-linear function approximation is preferred in the DSR-SMOO/CMOO MDP problems instead of other techniques due to the reduced complexity and due to their ability of obtaining a better prediction quality when compared with the linear function approximations. As discussed in Chapter 4, the MLPNN gets stuck in the local optimum and can affect the prediction quality (over-fitting) of the state or state-action values in the RL approach. In order to avoid these drawbacks, the number of layers N_L^{MLP} and the number of nodes for each layer N_N^l play a crucial role. Another factor which implies the over-fitting and the under-fitting problems is the features of the controller state space $\mathcal{S}_t^C$.

In DSR-SMOO/CMOO MDP problems, the scheduler controller may spend too much time in a given part of state $\mathcal{S}_t^C$ which can lead, in fact, to a poor generalization of the predicted values for some RL algorithms.

Based on the MLPNN function approximation from Eq. 5.17, the predicted controller state value $V_t^F\left(\mathcal{W}_t, \mathcal{S}_t^C\right)$ is entitled the MLPNN forwarded value of the controller state and can be expressed as:

$$V_t^F\left(\mathcal{W}_t, \mathcal{S}_t^C\right) = \mathcal{F}_{MLP}\left(\mathcal{W}_t, \mathcal{S}_t^C\right) \tag{5.18}$$

where the argument function is the controller state $\mathcal{S}_t^C$ and then, the stochastic target value of the controller state according to Eq. 5.12 becomes:

$$V_{t+1}^T\left(\mathcal{W}_t, \mathcal{S}_t^C\right) = \mathcal{R}\mathcal{W}_{t+1}\left(\mathcal{S}_t^C\right) + \gamma \cdot V_{t+1}^F\left(\mathcal{W}_t, \mathcal{S}_t^C\right) \tag{5.19}$$

Then, the estimated state value error can be rewritten such as:

$$\widetilde{E}_{t+1}^V\left(\mathcal{W}_t, \mathcal{S}_t^C\right) = \frac{1}{2}\left[V_{t+1}^T\left(\mathcal{W}_t, \mathcal{S}_t^C\right) - V_{t+1}^F\left(\mathcal{W}_t, \mathcal{S}_t^C\right)\right]^2 \tag{5.20}$$

The state-action values can be approximated in a similar way by using the MLPNN non-linear function approximation. The only difference reveals the fact that *one function approximation is used for each discrete action* $\mathcal{A}_t^a$. When the action set is continuous $\mathcal{A} \in \mathbb{R}^{D[\mathcal{A}]}$, one neural network is used to output the continuous action. Details about this concept are introduced in Sub-section 5.6.6. For the discrete case, it can be assumed that the MLPNN function approximation provides value for each discrete action in a given controller state $\mathcal{S}_t^C$. Thus, the predicted (forwarded) state-action value can be interpreted as shown by Eq. 5.21:

$$Q_t^F\left(\mathcal{W}_t^a, \mathcal{S}_t^C, \mathcal{A}_t^a\right) = \mathcal{F}_{MLP}^a\left(\mathcal{W}_t^a, \mathcal{S}_t^C\right) \tag{5.21}$$

where $\mathcal{W}_t^a = \left[\mathcal{W}_{t,1}^a, ..., \mathcal{W}_{t,N_L^{MLP}}^a\right]$ is the set of weights to be adjusted during the training stage for action $\mathcal{A}_t^a$ and $\mathcal{F}_{MLP}^a(\cdot)$ is the MLPNN function approximation for the same action. The stochastic target value for the state-action pair becomes:

$$Q_{t+1}^T\left(\mathcal{W}_t^a, \mathcal{S}_t^C, \mathcal{A}_t^a\right) = \mathcal{R}\mathcal{W}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) + \gamma \cdot \underset{a'}{max}\, Q_t^F\left(\mathcal{W}_t^a, \mathcal{S}_{t+1}^C, \mathcal{A}_{t+1}^{a'}\right) \tag{5.22}$$

where the argument is the controller state and then, the predicted squared error is:

$$\widetilde{E}_{t+1}^Q\left(\mathcal{W}_t^a, \mathcal{S}_t^C, \mathcal{A}_t^a\right) = \frac{1}{2}\left[Q_{t+1}^T\left(\mathcal{W}_t^a, \mathcal{S}_t^C, \mathcal{A}_t^a\right) - Q_{t+1}^F\left(\mathcal{W}_t^a, \mathcal{S}_t^C, \mathcal{A}_t^a\right)\right]^2 \tag{5.23}$$

The idea of the back-propagation is to adjust the sets of weights $\left(\mathcal{W}_t^{\cdot}, \mathcal{W}_t^a\right)_{a=1,..,|\mathcal{A}|}$ TTI-by-TTI in order to minimize as much as possible the predicted mean squared error for the state and the state-action values. As shown in Chapter 4, the back-propagation principle is based on the stochastic gradient descent method which is performed at each TTI. Let us define the MLPNN non-linear function definition domain for the state and state action values as: $\left(\mathcal{F}_{MLP}, \mathcal{F}_{MLP}^a\right): \mathbb{R}^{D[\mathcal{W}]} \times \mathbb{R}^{D\left[\mathcal{S}_t^C\right]} \to \mathbb{R}$ and the predicted error functions such as $\left(\widetilde{E}^Q, \widetilde{E}^V\right): \mathbb{R}^{D[\mathcal{W}]} \to \mathbb{R}$, where $D\left[\mathcal{W}\right]$ represents the total number of weights from each MLPNN structure. As shown already in Chapter 4, the gradient descent principle aims to update the parameters of $\left(\widetilde{E}^Q, \widetilde{E}^V\right)$ in the way that the update of each weight should lie in the direction of the negative gradient. Then, each weight can be updated based on Eq. 5.24:

$$\mathcal{W}_{t+1,w} = \mathcal{W}_{t,w} - \eta_t^V \cdot \frac{\partial \widetilde{E}_{t+1}^V\left(\mathcal{W}_{t,w}, \mathcal{S}_t^C\right)}{\partial \mathcal{W}_{t,w}} \tag{5.24}$$

where $w = 1,..,D\left[\mathcal{W}\right]$. For the presentation clarity, the MLPNN layer index is omitted in Eq. 5.24. For precise details about the implementation of the gradient descent principle in the RBFNN structures and about the error back-propagation technique for each node and for each layer, the reader should follow Figure 4.7 from Chapter 4. Similarly, the set of weights for the state-action error function can be determined as follows:

$$\mathcal{W}_{t+1,w}^a = \mathcal{W}_{t,w}^a - \eta_t^Q \cdot \frac{\partial \widetilde{E}_{t+1}^Q\left(\mathcal{W}_{t,w}^a, \mathcal{S}_t^C, \mathcal{A}_t^a\right)}{\partial \mathcal{W}_{t,w}^a} \tag{5.25}$$

By applying the stochastic gradient descent algorithm (observations are provided TTI-by-TTI directly from the LTE scheduler), this approach can offer poor convergence properties, it can get stuck in the local minima, and the trained MLPNN function can be the subject of the over-fitting symptom based on the training controller input state. Precisely, this means that if the controller spends a lot of time during the exploration stage in a given part of the state space, the neural network might forget how to handle the previously visited parts of the state

space. Some approaches are proposed in [231] to overcome these drawbacks. One way to mitigate or to eliminate the undesired effects of the MLPNN function approximations is to use the *stochastic iterative gradient descent learning* [232]. Precisely, the central controller has to store some previous observations (previous state, previous action, reward, current state, current action) and to train on these, as well on the new observations. This additional stage is entitled ***Experience Replay (ER)*** and can randomly provide some observations seen in the past. For implementation reasons, the ER stage should be performed between exploration and exploitation. More details are provided in Sub-section 5.6.7. Based on these principles, the global minimum in the MLPNN error is not guaranteed, but the convergence can be achieved by imposing the following conditions of the learning rates $\sum_{t=0}^{t\to\infty}\eta_t^V=\infty$, $\sum_{t=0}^{t\to\infty}\left(\eta_t^V\right)^2<\infty$, $\sum_{t=0}^{t\to\infty}\eta_t^Q=\infty$, $\sum_{t=0}^{t\to\infty}\left(\eta_t^Q\right)^2<\infty$, and then, conditions from Equations 5.26 and 5.27 are obtained:

$$\lim_{t\to\infty}\left(\widetilde{E}_{t+1}^V\left(\mathcal{W}_t,\mathcal{S}_t^C\right)\right)=\min_{\mathcal{W}}\left(\widetilde{E}_{t+1}^V\left(\mathcal{W}_t,\mathcal{S}_t^C\right)\right) \tag{5.26}$$

$$\lim_{t\to\infty}\left(\widetilde{E}_{t+1}^Q\left(\mathcal{W}_t^a,\mathcal{S}_t^C,\mathcal{A}_t^a\right)\right)=\min_{\mathcal{W}}\left(\widetilde{E}_{t+1}^Q\left(\mathcal{W}_t^a,\mathcal{S}_t^C,\mathcal{A}_t^a\right)\right) \tag{5.27}$$

In the MLPNN function approximations, there are several sensitive parameters that have to be tuned:

1. **The number of hidden nodes and hidden layers** determines how flexible the MLPNN function is. A higher number of hidden nodes implies a more accurate but slower learning process, and implicitly, lower updates. Of course, the number of nodes for a good enough approximation of the state value and the state-action value depends on the number of elements in the state space. When the MLPNN is too flexible, the risk of over-fitting the trained observations becomes higher. It is the typical case when the obtained function represents the generalization of the input controller state space, and also the noise which is present in the input state. On the other hand, a lower number of hidden nodes implies a poorer generalization, and in general, it can learn faster. Then, the MLPNN structure is not flexible and the trained data can be affected by the under-fitting symptom.

2. ***The learning rates*** have to be carefully tuned as they help to normalize the input of the neural network. For accuracy, *all elements involved in the controller input state space should be normalized* and the reward functions should be scaled. If the rewards are between $\mathcal{RW}_{min}$ and $\mathcal{RW}_{max}$ for a given discount factor γ, then the scaled reward function becomes:

$$\mathcal{RW}_{Scaled} = \frac{\mathcal{RW} \cdot (1-\gamma)}{\mathcal{RW}_{max} - \mathcal{RW}_{min}} \tag{5.28}$$

This normalization ensures that the approximated action values are always between -1 and 1. This scaling does not affect the order of the learned policy but the output of the neural network should be scaled back by using Equation 5.29:

$$Q^{F}_{t,Sc}\left(\mathcal{W}^{a}_{t}, \mathcal{S}^{C}_{t}, \mathcal{A}^{a}_{t}\right) = \frac{Q^{F}_{t}\left(\mathcal{W}^{a}_{t}, \mathcal{S}^{C}_{t}, \mathcal{A}^{a}_{t}\right) \cdot \left(\mathcal{RW}_{max} - \mathcal{RW}_{min}\right)}{(1-\gamma)} \tag{5.29}$$

Unfortunately, there is no model to determine the learning rates and the number of hidden nodes automatically, and the learning speed depends on these parameters. Typically, lower learning rates imply slower learning, but more accurate final results. Setting the learning rate too high can result in divergence of the network weights and nonsensical solutions (if using the standard Q-learning algorithm).

5.5.4 Policy Improvement and Policy Evaluation

The role of the reinforcement learning is to extract the optimal policy from the DSR-SMOO/CMOO MDP problems. In some situations, the RL approach may be interested in the policy value in order to *improve* the scheduling policy learned so far. Or in other circumstances, the RL algorithm may require the action extraction in order to *evaluate* the performance for a given learned policy. If the policy evaluation is performed TTI-by-TTI, then the obtained stage is entitled *exploitation (testing)*. If the policy is iterated TTI-by-TTI by using an improvement step or an evaluation step, the obtained stage is entitled *exploration (training)*. As mentioned earlier, the experience replay stage is required for some RL approaches in order to enhance the convergence speed and to avoid the local minima and the over-fitting problems when the MLPNN weights are trained. The

experience replay stage (stochastic exploration) uses the same steps when compared with the original exploration in terms of the policy evaluation and the policy improvement.

Let us define the set of stochastic policy values for a given controller state $\mathcal{S}_t^C$ at TTI t as: $\pi_t\left(\mathcal{S}_t^C\right)=\left\{\pi_{a,t}\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right)\right\}_{a=1,..,|\mathcal{A}|}$. Then, any improvement of these policies $\pi_{a,t}\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right), \forall a=1,..,|\mathcal{A}|$ is leading to the controller improved policy in state $\mathcal{S}_t^C$ such that $\pi_t^I\left(\mathcal{S}_t^C\right)$. Otherwise, the state policy remains in the original form expressed by $\pi_t^E\left(\mathcal{S}_t^C\right)$. The evaluation of the learned policy implies:

$$\pi_t^E\left(\mathcal{S}_t^C\right)\mapsto\left\{Q_t^{F,\pi^E}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right),V_t^{F,\pi^E}\left(\mathcal{S}_t^C\right)\right\},\ \forall a=\underset{a^*}{arg\,max}\left[Q_t^{F,\pi^E}\left(\mathcal{S}_t^C,\mathcal{A}_t^{a^*}\right)\right] \quad (5.30)$$

whereas the improvement step implies the random choice of actions as follows:

$$\left\{Q_t^{F,\pi^I}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right),V_t^{F,\pi^I}\left(\mathcal{S}_t^C\right)\right\}\mapsto\pi_t^I\left(\mathcal{S}_t^C\right),\ \forall a=1,...,|\mathcal{A}| \quad (5.31)$$

The scheduling policy $\pi_t^I\left(\mathcal{S}_t^C\right)$ is considered an improvement if and only if the obtained state and state-action values respect the conditions specified in Equations 5.32 and 5.33 such as:

$$Q_t^{F,\pi^I}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right)\geq Q_t^{F,\pi^E}\left(\mathcal{S}_t^C,\mathcal{A}_t^a\right),\ \forall a=1,...,|\mathcal{A}| \quad (5.32)$$

$$V_t^{F,\pi^I}\left(\mathcal{S}_t^C\right)\geq V_t^{F,\pi^E}\left(\mathcal{S}_t^C\right) \quad (5.33)$$

Then, the controller policy iteration can be achieved TTI-by-TTI until it reaches the optimal form, as shown in Eq. 5.34, with the premise that the DSR-SMOO/CMOO MDP problems are episodic:

$$\pi_0^E\mapsto\left\{Q_1^{F,\pi^E},V_1^{F,\pi^E}\right\}\mapsto\pi_2^I\mapsto\pi_3^I\mapsto....\left\{Q_t^{F,\pi^E},V_t^{F,\pi^E}\right\}\mapsto...\pi_{t+N_{TTI}^O}^*\mapsto\left\{Q_{t+N_{TTI}^O}^{F,\pi^*},V_{t+N_{TTI}^O}^{F,\pi^*}\right\} \quad (5.34)$$

Based on Eq. 5.34, the policy improvement and evaluation have to be combined from one TTI to another during the exploration stage. One way to achieve the tradeoff between improvement and evaluation is to use the parameter $\varepsilon\in[0,1]$

which permits to select random actions at each TTI (policy improvement) if $\varepsilon = 0$ or to follow the learned policy at each TTI (policy evaluation) if $\varepsilon = 1$. This exploration is entitled greedy exploration and permits to select greedy actions with a probability of $(1-\varepsilon)$. Then, the scheduling policy is considered to be ε-*greedy*. For the particular problem of the DSR-SMOO/CMOO MDP problem, the greedy parameter may differ depending on the type of RL approach.

For a given number of TTIs , the greedy parameter can be set based on the exponential function $\varepsilon_t = c_1 \cdot \exp(t / c_2)$ which permits to explore more at the beginning of the simulation and then, when approaching to the end of the scheduling simulation, the evaluation period takes more than the improvement period. This approach can be suitable for deterministic problems, and for these reasons, the above principle cannot obtain better policies when compared with the case when ε is constant during the whole scheduling simulation. The ε- *greedy* exploration can provide very good performance when the action state space is continuous (more details in Chapters 6 and 7) but it is not a great solution for the discrete action set when the discussed principle cannot detect what actions are good enough and which action should be improved by the greedy policy.

One way to avoid this drawback is to use the Boltzmann exploration. The Boltzmann principle aims to select those action values that present the highest probability distribution whereas the other actions may be ignored for a given input controller state space $\mathcal{S}_t^C$. The principle is shown in Eq. 5.35:

$$\pi_{a,t}^I\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right) = \exp\left(\frac{Q_t^F\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right)}{\tau}\right) \Bigg/ \sum_{a*=1}^{|\mathcal{A}|} \exp\left(\frac{Q_t^F\left(\mathcal{S}_t^C, \mathcal{A}_t^{a*}\right)}{\tau}\right) \qquad (5.35)$$

where τ is the temperature factor that sets how greedy the policy is. When $\tau \to \infty$, then the probability is $\pi_{a,t}^I\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right) \to 1$ and the policy becomes greedy (no evaluation) and when $\tau \to 0$, then $\pi_{a,t}^I\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right) \to 0$ (only evaluation) and the policy is more random [203]. If an action that has a higher state-action value when compared against other actions based on Eq. 5.35, then the probability of selecting that action becomes higher when compared with other discrete actions.

5.6 RL Algorithms in LTE Scheduling

In terms of the updating procedures for state and state-action values, the RL algorithms can take different forms. In many applications, different RL algorithms may behave differently. For this reason, different architectures of RL methods are applied for the DSR-SMOO/CMOO MDP problems in order to find the best set of scheduling policies. These techniques are not new, but for the DSR-SMOO/CMOO problems, the implementation of these algorithms represents part of the novelties proposed in this study.

In particular, the DSR-SMOO problem aims to find the optimal objective state by applying TTI-by-TTI different scheduling rules which follow the same target. For the QoS objectives, the MDP considers the state space to be continuous and the action set to be discrete. An interesting approach captures the attention for the fairness target. In this case, the PF scheduling rule has to be parameterized in such a way that at each TTI, the CDF of the normalized user throughput should lie in a given side of its domain. The parameterization can be achieved in two ways: by setting some fixed steps in the GPF parameters in order to permit the controller to move along the available state space or to use directly the output of the MLPNN function to set these steps continuously. In the first case, the action set is discrete whereas in the second one is continuous. In all cases, the controller architecture should be as simple as possible in order to face the time constraint imposed by the LTE scheduling procedure.

In Fig. 5.3 is highlighted the simplified architecture for the DSR-SMOO/CMOO MDP controller when the action state space is considered to be discrete. The role of the central controller can be divided into several tasks:

- Determines in the current TTI t if the learned policy should be evaluated or improved by using the exploration probabilities from Sub-section 5.5.4.
- A given set of actions, states and rewards are stored by this module in order to be used by the ER stage. The central controller is concentrated more on storing those actions when the reward value is $\mathcal{RW}_{t+1}\left(\mathcal{S}_t^C, \mathcal{A}_t^a\right) = 1$ and the rest of observations are randomly saved.

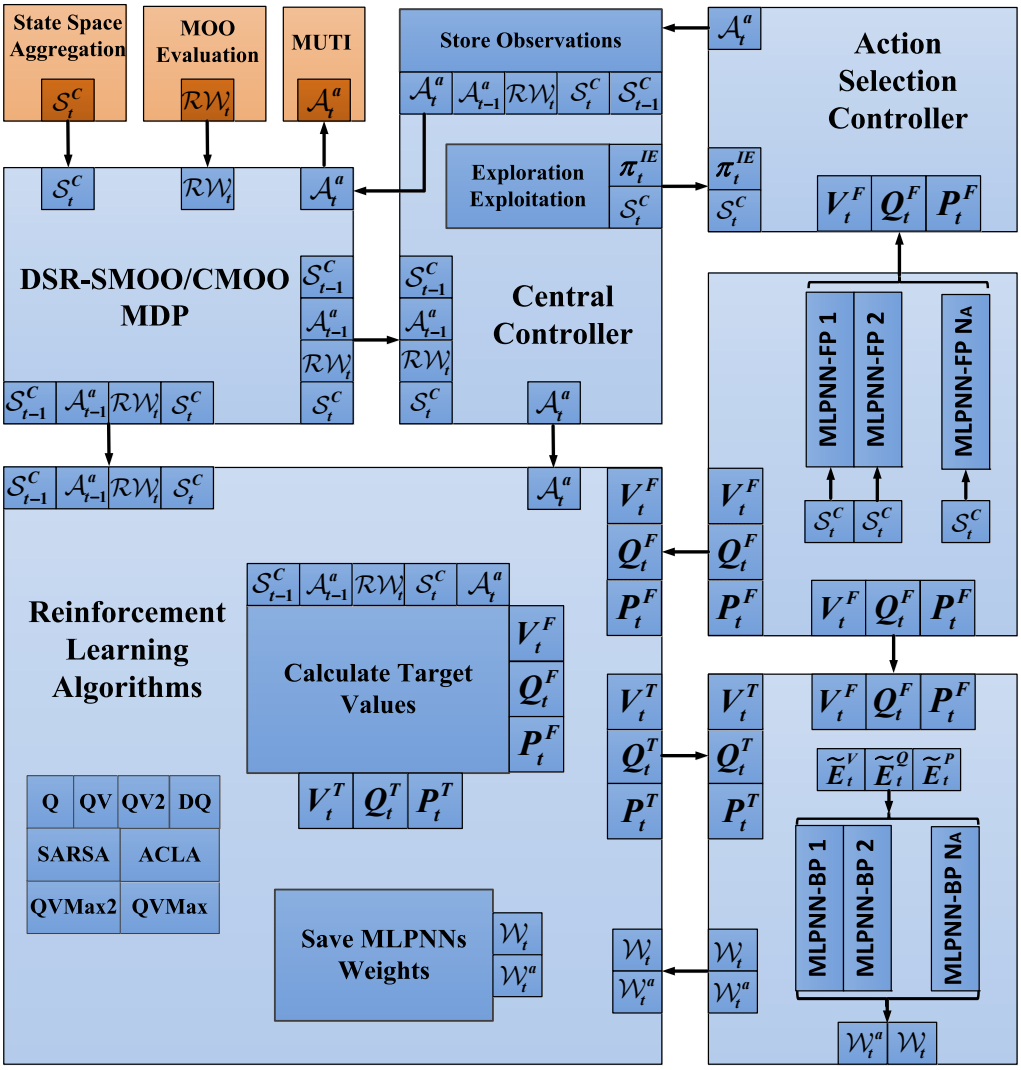


Fig. 5.3 The DSR-SMOO/CMOO MDP Controller

- The central controller is responsible also for choosing a proper RL approach which is used for a given packet scheduling session.

Each RL algorithm which is indicated in Fig. 5.3 will be described in the next sub-sections. The RL algorithms can take two main directions if the target state and the target state-action values follow or not the direction imposed by the learned scheduling policy such as:

- ***Off-policy RL approaches***: The updates do not follow the learned policy. For instance, in Eq. 5.22, the target state-action value $Q_{t+1}^{T}\left(\mathcal{W}_t^a, \mathcal{S}_t^C, \mathcal{A}_t^a\right)$ is calculated based on the maximization procedure of other actions which in fact do not follow the policy $\pi_{a,t}\left(\mathcal{A}_t^a \mid \mathcal{S}_t^C\right)$ learned so far.

- ***On-policy RL approaches***: The updates follow the learned policies. Most of the on-policy RL approaches provide a better performance for the DSR-SMOO MDP problems when compared against the typical off-policy approaches. For the NGMN fairness requirement, the proposed sustainable scheduling policies being trained by using the on-policy RL algorithms provide the best results in terms of the number of feasible TTIs. More details about these results are presented in Chapter 6.

The action selection controller is responsible mainly for choosing the best controller action based on state and state-action values provided by the MLPNN function approximations through the feed-forward procedure. When the policy improvement stage is considered, then the controller selects that action with the highest probability distribution. When the policy evaluation step is performed, the action value with the highest MLPNN approximation is selected based on the instantaneous controller state. For convenience, in the updating formulas, TTI t will be referred as a current time instant when the packet scheduling is performed, rather that TTI $t+1$ which was used in the previous sub-sections.

In the following sub-sections, the existing RL algorithms are extracted from the specialty literature and applied to the controlling process of LTE/LTE-A scheduling. The rest of this section is organized as follows: Sub-section 5.6.1 presents the Q-learning algorithm (off-policy) [228], Sub-section 5.6.2 extends the Q-learning approach to Double-Q-learning (off-policy) [234] and Sub-section 5.6.3 introduces the principle of SARSA-learning [235] which is an on-policy RL algorithm. Sub-section 5.6.4 presents the state-value RL algorithms such as [238]: QV (on-policy), QV2 (on-policy), QVMAX (off-policy) and QVMAX2 (off-policy). The Actor Critic Learning Automata (ACLA) [239] is presented in Sub-section 5.6.5 as an actor-critic and on-policy RL scheme. These RL techniques use discrete action spaces. The Continuous ACLA (CACLA) [240] is an on-policy method which makes use of continuous action space as shown in Sub-section 5.6.6. The sustainable sets of scheduling policies are obtained based on the RL approaches introduced above. Chapter 6 presents the performance of scheduling policies for the DSR-SMOO problems and Chapter 7 analyzes the advantages of using the set of sustainable policies for the DSR-CMOO problems.

5.6.1 Q-Learning

The goal of the scheduler controller is to learn the optimal policy $\pi_{\mathcal{A}}^{*}$ that maximizes the expected accumulated reward starting from any initial state. The reward function and the transition probabilities are not known in advance. Due to the fact that there is no model of the RRM environment for different scheduling rules, the policy of actions must be learned based on the multiple trials and errors between the RRM environment (including the packet scheduler) and the scheduler controller. The Q-learning algorithm is the best known RL technique which is used in the control systems for many years [228], [233]. It is considered to be an off-policy learning procedure since its update aims to include the maximum state-action value. The one step TTI Q-learning approximation is defined by Eq. 5.36:

$$Q_t^T\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)\xleftarrow{\eta^Q}\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)+\gamma\cdot\max_{a'}Q_t^F\left(\mathcal{S}_t^C,\mathcal{A}_t^{a'}\right) \qquad (5.36)$$

The error calculation is based on Eq. 5.23 and is reloaded in Eq. 5.37 at the current scheduling time instant TTI *t*:

$$\widetilde{E}_t^Q\left(\mathcal{W}_{t-1}^a,\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)=\frac{1}{2}\left[Q_t^T\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)-Q_t^F\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)\right]^2 \qquad (5.37)$$

The forwarded action-value function $Q_t^F\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)$ can approximate the optimal state-action value Q^* regardless to the policy which has been learned so far. The error in Eq. 5.37 is back-propagated in order to apply the weight correction to the MLPNN structure. In Fig. 5.4 it is shown the basic principle of the Q-learning algorithm applied for the DSR-SMOO/CMOO problems. Algorithm 5.1 shows the main steps involved in Q-learning for the DSR-SMOO/CMOO MDP problems.

5.6.2 Double Q-Learning

The classical Q-learning algorithm provides a poor performance due to the capacity of overestimating for the state-action values [234]. The Double Q (DQ)-learning is proposed to underestimate the action values rather than to overestimate them [234]. The DQ-learning is obtained by using a double predicted action

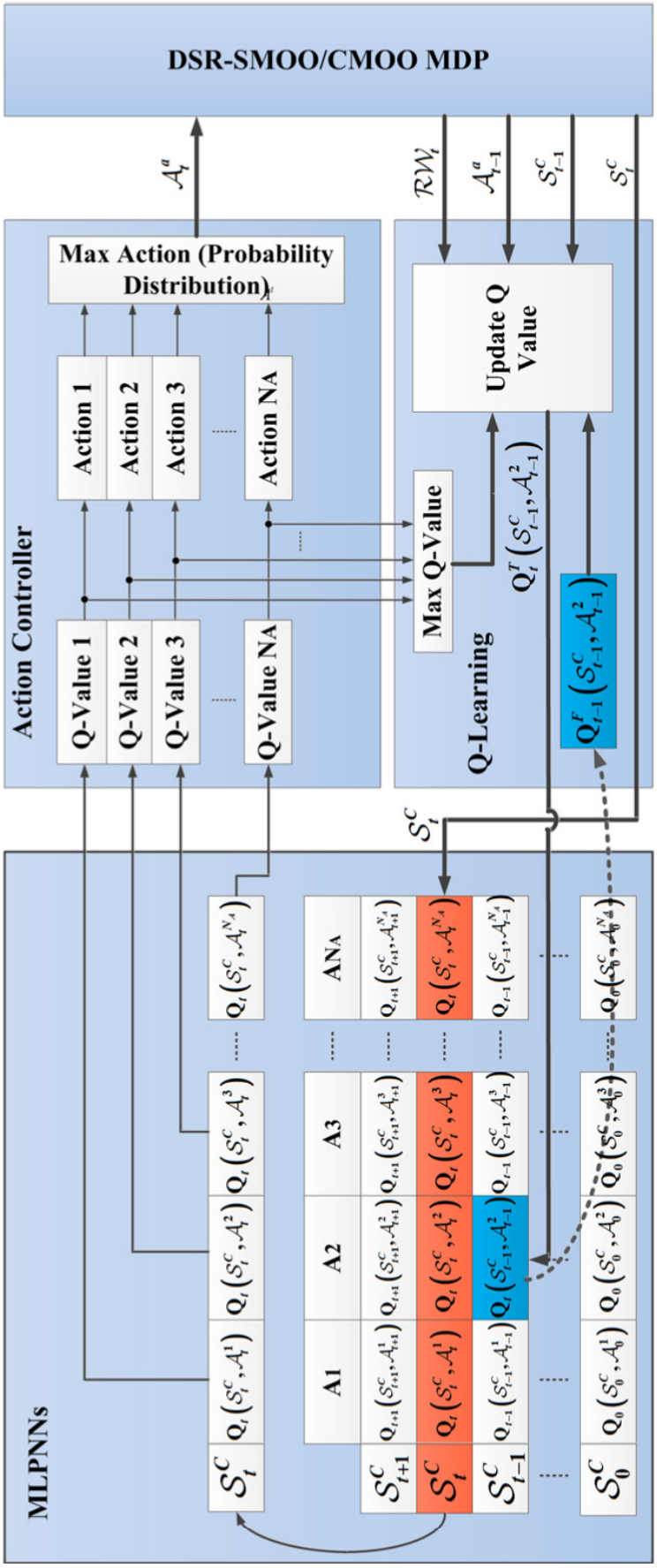


Fig. 5.4 Q-Learning Based DSR-SMOO/CMOO

Algorithm 5.1 Q-Learning Based DSR-SMOO/CMOO MDP

1. **for** each TTI t
2. Scheduler State Space Aggregation: $\mathcal{S}_t^S \xrightarrow{Agg.} \mathcal{S}_t^C$
3. MOO Evaluation procedure: $\mathcal{RW}_t \leftarrow \mathcal{F}\left(\mathcal{S}_{t-1}^C.\mathcal{A}_{t-1}^a\right)$
4. Verify if $\mathcal{S}_t^C$ is terminal $\left(\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)=1\right)$
5. Explore $\mathcal{S}_t^C$ for all actions $\mathcal{A}_t^{a'}, a'=1,..,|\mathcal{A}|$ based on $\pi_t^{IE}\left(\mathcal{A}_t^{a'} \mid \mathcal{S}_t^C\right)$
6. **if** is policy evaluation (MLPNN FP): $\mathcal{A}_t^a = \arg\max_{a'}\left(\mathcal{F}_{MLP}^{a'}\left(\mathcal{S}_t^C,\mathcal{A}_t^{a'}\right)\right)$
7. **else** is policy improvement: $\mathcal{A}_t^a = \arg\max_{a'}\left(\pi_t^I\left(\mathcal{A}_t^{a'} \mid \mathcal{S}_t^C\right)\right)$
8. **if** $\left(\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)=1\right)$ store observations $\left(\mathcal{S}_{t-1}^C.\mathcal{A}_{t-1}^a,\mathcal{RW}_t,\mathcal{S}_t^C,\mathcal{A}_t^a\right)$
9. **else if** $(1-\varepsilon)<\varepsilon_{INIT}$ store observations $\left(\mathcal{S}_{t-1}^C.\mathcal{A}_{t-1}^a,\mathcal{RW}_t,\mathcal{S}_t^C,\mathcal{A}_t^a\right)$
10. **Update** $Q_t^T\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right) \xleftarrow{\eta^Q} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)+\gamma\cdot\max_{a'} Q_t^F\left(\mathcal{S}_t^C,\mathcal{A}_t^{a'}\right)$
11. **Back Propagate Error (MLPNN BP):** $\widetilde{E}_t^Q\left(\mathcal{W}_{t-1}^a,\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)$
12. **Update** MLPNN weights: $\mathcal{W}_{t,w}^a = \mathcal{W}_{t-1,w}^a - \eta_t^Q\cdot\partial\widetilde{E}_t^Q\left(\mathcal{W}_{t-1,w}^a,\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)/\partial\mathcal{W}_{t-1,w}^a$
13. **MUTI: map** $\mathcal{A}_t^a$ **to scheduling rule** $c_{o,w_o}[t]$
14. $c_{o,w_o}[t] \rightarrow$ **calculate metrics, RB allocations and TBS calculation**
15. **end** TTI t

values by keeping the same characteristics of the off-policy RL. In other words, the DQ-learning is in fact a RL multi-agent system in which the action value of Agent A is used in the target action values of Agent B, and the value of Agent B is used to compute the target action values of Agent A. The resulted policies are combined by using some ensemble algorithms in order to provide the final policy [203]. For the current purpose of the DSR-SMOO/CMOO MDP problems, the DQ-learning implies two action values $Q_t^{AT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)$ and $Q_t^{BT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)$ which are updated according to Eq. 5.38.a) and Eq. 5.38.b), respectively:

$$Q_t^{AT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right) \xleftarrow{\eta^Q} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)+\gamma\cdot Q_t^{BF}\left(\mathcal{S}_t^C,\mathcal{A}_t^{a^*}\right) \tag{5.38.a}$$

$$Q_t^{BT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right) \xleftarrow{\eta^Q} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)+\gamma\cdot Q_t^{AF}\left(\mathcal{S}_t^C,\mathcal{A}_t^{b^*}\right) \tag{5.38.b}$$

where $\mathcal{A}_t^{a^*} = \underset{a'}{arg\,max}\, Q_t^{AT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^{a'}\right)$ and $\mathcal{A}_t^{b^*} = \underset{b'}{arg\,max}\, Q_t^{BT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^{b'}\right)$. The third Agent C can combine both policies by simply averaging its state-action target

such as: $Q_t^{CT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)=\left[Q_t^{AT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)+Q_t^{BT}\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)\right]/2$. This way, DQ-learning makes use of both agent experiences. If the overestimation can be mitigated by using the DQ-learning, the ER stage is still needed for this approach. In the exploitation stage, only Agent C is used in the policy exploitation. For some multi-objective performance metrics, the DQ policy can outperform other RL policies (to be detailed in Chapter 7).

5.6.3 SARSA Learning

If the exploration effect is totally considered for the state-action updates, then the considered RL algorithm is considered to be on-policy. In other words, the selection of the current action depends only on the probability distribution of each action $\mathcal{A}_t^a$ for a given controller state $\mathcal{S}_t^C$. It is the case of SARSA learning [235],[236] which updates the state-action values based on the MDP problem $\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a,\mathcal{RW}_t,\mathcal{S}_t^C,\mathcal{A}_t^{a'}\right)$ where $(a,a')=1,..,|\mathcal{A}|$. In this sense, the action $\mathcal{A}_t^{a'}$ follows the current behavior of policy $\pi_{\mathcal{A},t}$ for a given set of actions. Thus, the target state-action value is calculated according to Eq. 5.39 and the error function keeps a similar form to Eq. 5.37.

$$Q_t^T\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)\xleftarrow{\eta^Q}\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_{t-1}^a\right)+\gamma\cdot Q_t^F\left(\mathcal{S}_t^C,\mathcal{A}_t^{a'}\right) \tag{5.39}$$

The principles of SARSA learning are shown in Fig. 5.5 where the MLPNN approximations are highlighted as look-up tables for the visited state-action pairs for a more comprehensive explanation. It can be seen that the max operator is not included and the selected action $\mathcal{A}_t^3$ is considered by the Q value updating block as a new action at TTI *t*. SARSA is considered to converge to the optimal policy when the state-action pairs are visited for infinite number of steps [42]. Details about the implementation of SARSA-learning for the DSR-SMOO/CMOO MDP problems are given in Algorithm 5.2. Lines 7 and 8 save the visited observation if the current reward is 1, or in the opposite case, the greedy action decides whether or not the observation deserves to be saved for the ER stage. SARSA shows very good performances in Chapter 7 when solves the DSR-CMOO MDP problems.

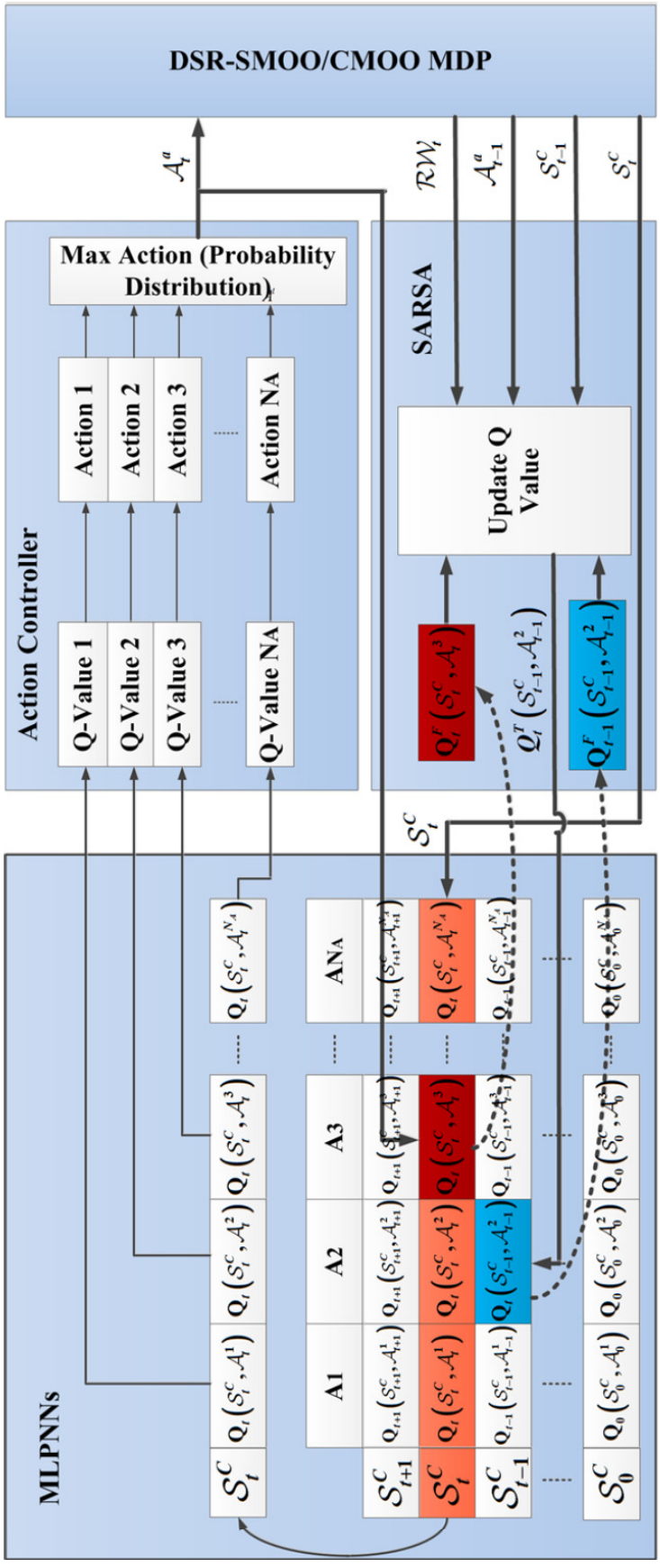


Fig. 5.5 SARSA-Learning Based DSR-SMOO/CMOO MDP

Algorithm 5.2 SARSA-Learning Based DSR-SMOO/CMOO MDP

1. **for** each TTI t
2. **Scheduler State Space Aggregation**: $\mathcal{S}_t^S \xrightarrow{Agg.} \mathcal{S}_t^C$
3. **MOO Evaluation procedure**: $\mathcal{RW}_t \leftarrow \mathcal{F}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$
4. **verify** if $\mathcal{S}_t^C$ is terminal $\left(\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)=1\right)$
5. **explore** $\mathcal{S}_t^C$ for all actions $\mathcal{A}_t^b, b=1,..,|\mathcal{A}|$ based on $\pi_t^{IE}\left(\mathcal{A}_t^b \mid \mathcal{S}_t^C\right)$
6. **determine** $\mathcal{A}_t^{a'}$ based on Eq. 5.35 (Boltzmann Distribution)
7. **if** $\left(\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C.\mathcal{A}_{t-1}^a\right)=1\right)$ store observations $\left(\mathcal{S}_{t-1}^C.\mathcal{A}_{t-1}^a, \mathcal{RW}_t, \mathcal{S}_t^C, \mathcal{A}_t^{a'}\right)$
8. **else if** $(1-\varepsilon)<\varepsilon_{INIT}$ store observations $\left(\mathcal{S}_{t-1}^C.\mathcal{A}_{t-1}^a, \mathcal{RW}_t, \mathcal{S}_t^C, \mathcal{A}_t^{a'}\right)$
9. **update** $Q_t^T\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) \xleftarrow{\eta^Q} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)+\gamma \cdot Q_t^F\left(\mathcal{S}_t^C, \mathcal{A}_t^{a'}\right)$
10. **back-propagate error (MLPNN BP):** $\widetilde{E}_t^Q\left(\mathcal{W}_{t-1}^a, \mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$
11. **update** weights**:** $\mathcal{W}_{t,w}^a=\mathcal{W}_{t-1,w}^a-\eta_t^Q \cdot \partial \widetilde{E}_t^Q\left(\mathcal{W}_{t-1,w}^a, \mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) / \partial \mathcal{W}_{t-1,w}^a$
12. **MUTI: map** $\mathcal{A}_t^{a'}$ **to scheduling rule** $c_{o,w_o}[t]$
13. $c_{o,w_o}[t] \rightarrow$ **calculate metrics, RB allocations and TBS calculation**
14. **end for**

5.6.4 QV-Learning

The RL algorithms discussed so far aim to consider only the state-action pairs for the policy improvement and evaluation. Another common approach is to combine the information of the state and the state-action values in order to build the optimal policy. QV-learning is one of these algorithms, which is very similar to SARSA but considers the step updates such that [237]:

$$Q_t^T\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) \xleftarrow{\eta^Q} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)+\gamma \cdot V_t^F\left(\mathcal{S}_t^C\right) \tag{5.40.a}$$

$$V_t^T\left(\mathcal{S}_{t-1}^C\right) \xleftarrow{\eta^V} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C\right)+\gamma \cdot V_t^F\left(\mathcal{S}_t^C\right) \tag{5.40.b}$$

The error functions that are back-propagated TTI-by-TTI are calculated with the respect of Equations 5.20 and 5.23. QV-learning is considered an on-policy RL approach since it follows the learned policy. For some DSR-SMOO/CMOO MDP problems, the state value can approximate better than other state-action values and thus can improve the prediction of state-action values. The main principles of QV-

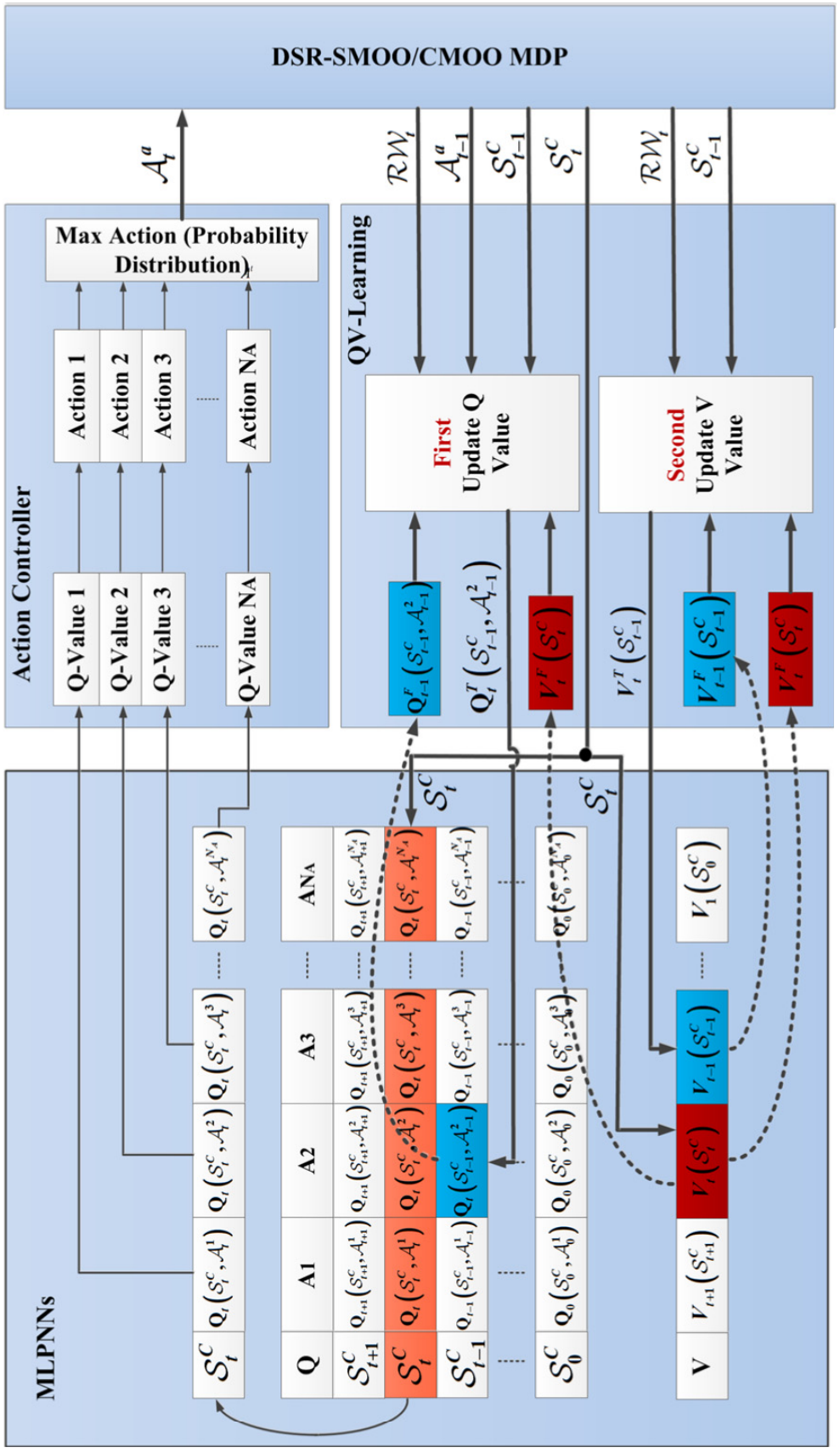


Fig. 5.6 QV-Learning Based DSR-SMOO/CMOO MDP

Algorithm 5.3 QV-Learning Based DSR-SMOO/CMOO MDP

1. for each TTI t

2. **Scheduler State Space Aggregation**: $\mathcal{S}_t^S \xrightarrow{Agg.} \mathcal{S}_t^C$

3. **MOO Evaluation procedure**: $\mathcal{RW}_t \leftarrow \mathcal{F}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$

4. **verify** if $\mathcal{S}_t^C$ is terminal $\left(\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a\right) = 1\right)$

5. **explore** $\mathcal{S}_t^C$ for all actions $\mathcal{A}_t^{a'}$, $a' = 1,..,|\mathcal{A}|$ based on $\pi_t^{IE}\left(\mathcal{A}_t^{a'} \mid \mathcal{S}_t^C\right)$

6. **if** is policy evaluation (MLPNN FP): $\mathcal{A}_t^a = \arg\max_{a'}\left(\mathcal{F}_{MLP}^{a'}\left(\mathcal{S}_t^C . \mathcal{A}_t^{a'}\right)\right)$

7. **else** is policy improvement: $\mathcal{A}_t^a = \arg\max_{a'}\left(\pi_t^I\left(\mathcal{A}_t^{a'} \mid \mathcal{S}_t^C\right)\right)$

8. **if** $\left(\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a\right) = 1\right)$ store observations $\left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a, \mathcal{RW}_t, \mathcal{S}_t^C, \mathcal{A}_t^a\right)$

9. **else if** $(1-\varepsilon) < \varepsilon_{INIT}$ store observations $\left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a, \mathcal{RW}_t, \mathcal{S}_t^C, \mathcal{A}_t^a\right)$

10. **feed-forward** $V_t^F\left(\mathcal{S}_t^C\right) = \mathcal{F}_{MLP}\left(\mathcal{S}_t^C\right)$

11. **update** $V_t^T\left(\mathcal{S}_{t-1}^C\right) \xleftarrow{\eta^V} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C\right) + \gamma \cdot V_t^F\left(\mathcal{S}_t^C\right)$

12. **back-propagate error:** $\widetilde{E}_t^V\left(\mathcal{W}_{t-1}, \mathcal{S}_{t-1}^C\right)$

13. **update** weights**:** $\mathcal{W}_{t,w} = \mathcal{W}_{t-1,w} - \eta_t^V \cdot \partial \widetilde{E}_t^V\left(\mathcal{W}_{t-1,w}, \mathcal{S}_{t-1}^C\right) / \partial \mathcal{W}_{t-1,w}$

14. **update** $Q_t^T\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) \xleftarrow{\eta^Q} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) + \gamma \cdot V_t^F\left(\mathcal{S}_t^C\right)$

15. **back-propagate error:** $\widetilde{E}_t^Q\left(\mathcal{W}_{t-1}^a, \mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$

16. **update** weights**:** $\mathcal{W}_{t,w}^a = \mathcal{W}_{t-1,w}^a - \eta_t^Q \cdot \partial \widetilde{E}_t^Q\left(\mathcal{W}_{t-1,w}^a, \mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) / \partial \mathcal{W}_{t-1,w}^a$

17. **MUTI: map** $\mathcal{A}_t^a$ **to scheduling rule** $c_{o,w_o}[t]$

18. $c_{o,w_o}[t] \rightarrow$ **calculate metrics, RB allocations and TBS calculation**

19. **end** TTI t

learning are shown in Fig. 5.6, and Algorithm 5.3 highlights the main steps involved for the DSR-SMOO/CMOO MDP purpose.

Other RL algorithms can be developed based on the combinations of state and state-action values for the updating equations [238]. One of these algorithms is ***QV2-learning*** which keeps the same form of targeting values as shown in Eq. 5.40.a with the only difference in the state value error which has to be back-propagated such that:

$$(QV2): \ \widetilde{E}_t^V\left(\mathcal{W}_{t-1}, \mathcal{S}_{t-1}^C\right) = \frac{1}{2}\left[V_t^T\left(\mathcal{S}_{t-1}^C\right) - Q_t^F\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)\right]^2 \qquad (5.41)$$

The QV2 error for the state value depends on the state-action value in the previous state and on the current state value. ***QVMAX-learning*** also keeps track of both values by modifying the updating rule for the state value and keeping other update and error calculation similar to the QV-learning case. The target state value calculation is based on Eq. 5.42 which reveals in fact the off-policy character of the updating rule.

$$(QVMAX):\ V_t^T\left(\mathcal{S}_{t-1}^C\right)\xleftarrow{\eta^V}\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C,\mathcal{A}_t^a\right)+\gamma\cdot\max_{a'}Q_t^F\left(\mathcal{S}_t^C,\mathcal{A}_t^{a'}\right) \quad (5.42)$$

QVMAX2-learning is practically a combination of QV, QV2 and QVMAX procedures. The state-action value is updated according to Eq. 5.40.a, and the error propagation is performed similarly to QV-learning. The state value is updated similarly to QVMAX (Eq. 5.42), and the error calculation is performed according to Eq. 5.23 which is similar to the Q-learning algorithm. QVMAX2-learning is an off-policy method since the max operator is involved in the updating procedures.

5.6.5 Actor Critic Learning Automata (ACLA)

The RL algorithms analyzed so far aim to improve or to evaluate the learned policy TTI-by-TTI. In other words, all these techniques behave as actors. One major drawback of these approaches refers to the fact that there is no online evaluation of how good the learned policy is. Therefore, a new entity which is able to criticize the applied action can improve and can speed-up the convergence to the optimal policy. QV-learning approaches can be seen as actor-critic schemes since the state values are used for the state-action updates. But they do not provide any explicit information if the applied action can lead the system in a desired feasible or optimal state. When the DSR-SMOO MDP problem focusing on the fairness performance is considered, the reward function provides the information about how far or close the system is from the optimal state. But the reward is not always enough to train the MLPNN weights in the direction of the optimal state. Then, a critic which informs the MLPNN structure if the current state value is better than the previous one can help improve the learning procedure in the

direction of the optimal scheduler state. For the current purpose of the DSR-SMOO/CMOO MDP problems, the required RL schemes perform as an actor when applying the learned policy of scheduling rules and as a critic when the applied action is criticized based on the state value.

Actor Critic Learning Automata (ACLA) is one of the most powerful actor-critic RLs proposed initially in [239]. The same principle of ACLA is applied for the DSR-SMOO/CMOO MDP problems. At each TTI, the state value is updated, and the back-propagation error between the target and the forwarded values is calculated. If the error is positive, this means that the previous action was a good choice and the probability of selecting that action should be increased. If the error is negative, then the probability of selecting that action for a given state value is decreased. The state-action values are considered here to be preference values $P_t\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a}\right)$ instead of quality values $Q_t\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a}\right)$ used for other RL approaches. The preference target value for the DSR-SMOO/CMOO MDP problems is calculated according to Eq. 5.43:

$$P_t^T\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a}\right)\xleftarrow{\eta^P}\mathcal{P}\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a}\right)-P_t^F\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a}\right) \tag{5.43}$$

where P_t^T, P_t^F are the target and the forwarded preference values, respectively, and $\mathcal{P}\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a}\right)$ is the critic decision which is calculated based on Eq. 5.44:

$$\mathcal{P}\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a}\right)=\begin{cases}-1, & \widetilde{E}_t^V\left(\mathcal{W}_{t-1},\mathcal{S}_{t-1}^{C}\right)<0\\ 1, & \widetilde{E}_t^V\left(\mathcal{W}_{t-1},\mathcal{S}_{t-1}^{C}\right)>0\end{cases} \tag{5.44}$$

The state values and the state error values are calculated similarly to the QV-learning procedure. In this case, Equation 5.44 is a critic whereas the probability distribution of each action depends on Eq. 5.43 when the Boltzmann distribution is used with the respect of the preference values. Even if ACLA is considered sub-optimal for particular RL problems [203], for many DSR-SMOO/CMOO MDP problems, the current approach is able to offer a better performance when compared with the previous versions of RL algorithms. Figure 5.7 shows the basic principles of the ACLA learning algorithm where the first and the third blocks are considered to be actors whereas the second one criticizes the performance of the

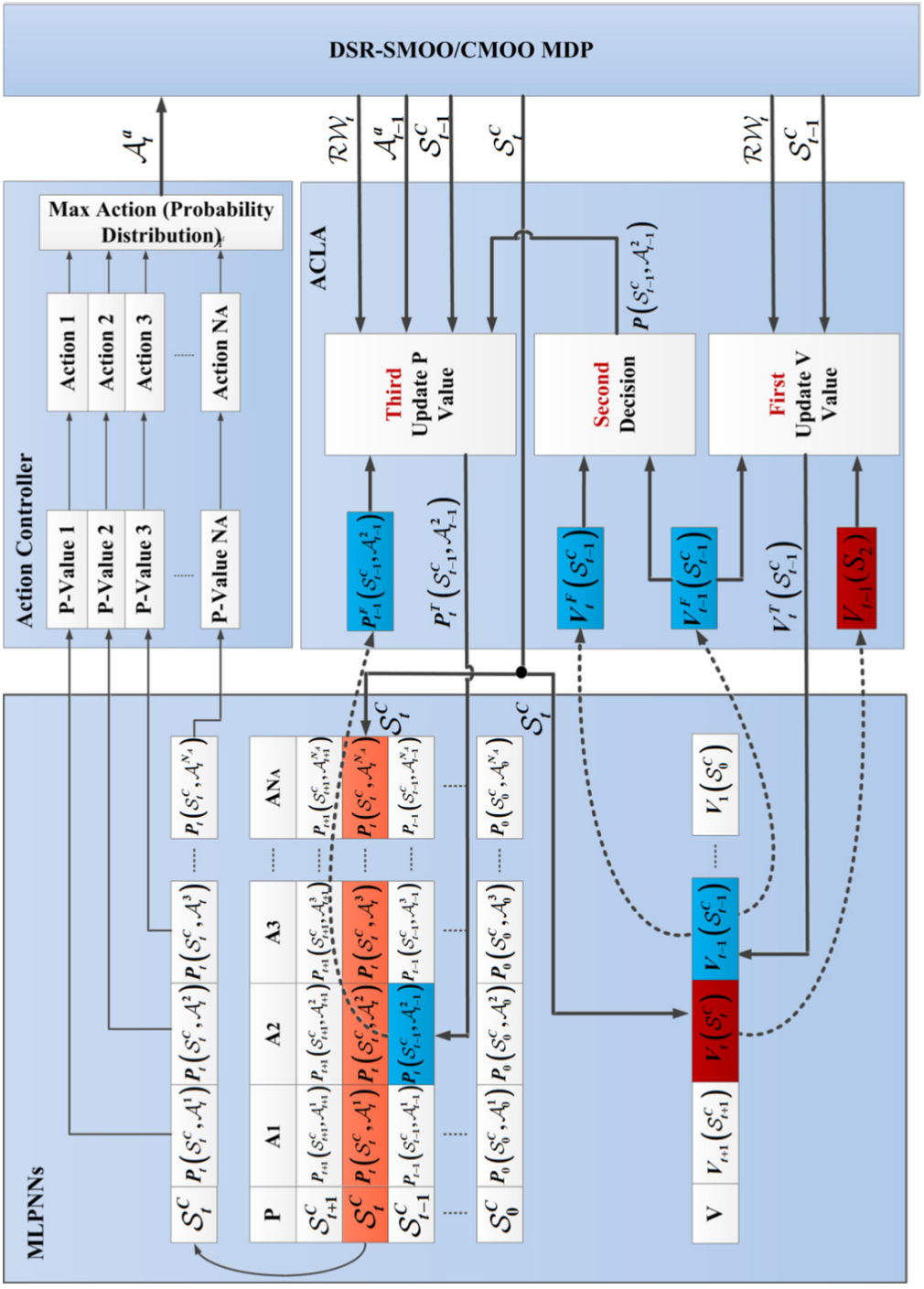


Fig. 5.7 ACLA-Learning Based DSR-SMOO/CMOO MDP

Algorithm 5.4 ACLA-Learning Based DSR-SMOO/CMOO MDP

1. for each TTI t
2. **Scheduler State Space Aggregation**: $\mathcal{S}_t^S \xrightarrow{Agg.} \mathcal{S}_t^C$
3. **MOO Evaluation procedure**: $\mathcal{RW}_t \leftarrow \mathcal{F}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$
4. **verify** if $\mathcal{S}_t^C$ is terminal $\left(\mathcal{RW}_t\ \left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a\right) = 1\right)$
5. **explore** $\mathcal{S}_t^C$ for all actions $\mathcal{A}_t^{a'}, a' = 1,..,|\mathcal{A}|$ based on $\pi_t^{IE}\left(\mathcal{A}_t^{a'} \mid \mathcal{S}_t^C\right)$
6. **if** is policy evaluation (MLPNN FP): $\mathcal{A}_t^a = \arg\max_{a'}\left(\mathcal{F}_{MLP}^{a'}\left(\mathcal{S}_t^C, \mathcal{A}_t^{a'}\right)\right)$
7. **else** is policy improvement: $\mathcal{A}_t^a = \arg\max_{a'}\left(\pi_t^I\left(\mathcal{A}_t^{a'} \mid \mathcal{S}_t^C\right)\right)$
8. **if** $\left(\mathcal{RW}_t\ \left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a\right) = 1\right)$ store observations $\left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a, \mathcal{RW}_t\ , \mathcal{S}_t^C, \mathcal{A}_t^a\right)$
9. **else if** $(1-\varepsilon) < \varepsilon_{INIT}$ store observations $\left(\mathcal{S}_{t-1}^C . \mathcal{A}_{t-1}^a, \mathcal{RW}_t\ , \mathcal{S}_t^C, \mathcal{A}_t^a\right)$
10. **feed-forward** $V_t^F\left(\mathcal{S}_t^C\right) = \mathcal{F}_{MLP}\left(\mathcal{S}_t^C\right)$
11. **update** $V_t^T\left(\mathcal{S}_{t-1}^C\right) \xleftarrow{\eta^V} \mathcal{RW}_t\ \left(\mathcal{S}_{t-1}^C\right) + \gamma \cdot V_t^F\left(\mathcal{S}_t^C\right)$
12. **back-propagate error :** $\widetilde{E}_t^V\left(\mathcal{W}_{t-1}, \mathcal{S}_{t-1}^C\right)$
13. **update** weights**:** $\mathcal{W}_{t,w} = \mathcal{W}_{t-1,w} - \eta_t^V \cdot \partial \widetilde{E}_t^V\left(\mathcal{W}_{t-1,w}, \mathcal{S}_{t-1}^C\right) / \partial \mathcal{W}_{t-1,w}$
14. **determine the critic decision** $\mathcal{P}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$ **based on Eq. 5.43**
15. **feed-forward** $P_t^F\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) = \mathcal{F}_{MLP}^a\left(\mathcal{S}_{t-1}^C\right), \forall a \in \mathcal{A}$
16. **update** $P_t^T\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) \xleftarrow{\eta^Q} \mathcal{P}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) - P_t^F\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$
17. **back propagate error:** $\widetilde{E}_t^P\left(\mathcal{W}_{t-1}^a, \mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$
18. **update** weights**:** $\mathcal{W}_{t,w}^a = \mathcal{W}_{t-1,w}^a - \eta_t^P \cdot \partial \widetilde{E}_t^P\left(\mathcal{W}_{t-1,w}^a, \mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) / \partial \mathcal{W}_{t-1,w}^a$
19. **MUTI: map** $\mathcal{A}_t^a$ **to scheduling rule** $c_{o,w_o}[t]$
20. $c_{o,w_o}[t] \rightarrow$ **calculate metrics, RB allocations and TBS calculation**
21. **end for**

current action. The implementation details for the DSR-SMOO/CMOO MDP problems with ACLA learning are provided in Algorithm 5.4 where the critic decision is pointed by Line 14. Once the critic decision and the forwarded value $P_t^F\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$ are determined, the error is back-propagated for the selected MLPNN function. In Chapter 6, ACLA learning provides the sustainable scheduling policies for the DSR-SMOO problems focusing on the NGMN fairness and GBR objectives, whereas in Chapter 7, the obtained policies show a very good sustainability when the NGMN fairness, GBR, HoL delay and PDR objectives are considered in solving the DSR-CMOO combinatorial problems.

5.6.6 Continuous ACLA (CACLA)

The RL approaches discussed and proposed so far aim to use a discrete set of actions which are very suitable for the LTE scheduling control. As mentioned earlier, for this purpose, most of the RL algorithms have the advantage of a very good scalability. In some cases, not only the scheduling rule index is required. For instance, the GPF scheduling rule offers the possibility of adjusting its parameters by improving or degrading the system throughput depending on different circumstances. Therefore, the DSR-SMOO MDP problems focusing on the fairness performance require to fine tune the GPF parameters in order to obtain a certain level of the fairness-throughput tradeoff at each TTI. Of course, this aspect can be achieved by using a finite number of parameter steps and to use one of the presented RL algorithms which performs the best. Unfortunately, the performance of these approaches depends on the parameter step lengths which have to be decided (details in Chapter 6). In fact, this leads to the incapability of the policy to adapt to very new circumstances (fluctuations in the number of active bearers).

Based on the above considerations, the continuous and possible multi-dimensional action space is required in order to control the parameterized functions involved in the scheduling procedure. The standard continuous ACLA (CACLA) proposed in [240] is very suitable for the current problem. CACLA is the continuous version of ACLA in which the action space is continuous rather than discrete (adapts the continuous fairness parameters α and β from Eq. 3.68).

CACLA keeps the same track as other RL algorithms in the sense that at each TTI t one of the exposed exploration methods (greedy or Boltzmann) is used in order to determine the action to be performed. By using the TD principle, the action is evaluated on whether or not was a good option to be applied in the previous TTI. The critic is performed here in terms of the state value calculation. If the critic error is positive, then the applied action was a good idea and it should be reinforced by back-propagating the actor error.

Let us define the continuous action $\mathcal{A}_t \in \mathbb{R}^{D[\mathcal{A}]}$, where $D[\mathcal{A}]$ represents the action space dimension. At the beginning of TTI t after the aggregation

procedure, the newest state $\mathcal{S}_t^C$ is sensed. Based on the CACLA principle, two MLPNN functions are used for the state-action and state values. Based on these principles, the action value $\mathcal{A}_t = A_t^F\left(\mathcal{S}_t^C\right)$ is reinforced based on Eq. 5.45:

$$A_t^T\left(\mathcal{S}_{t-1}^C\right) \leftarrow A_t^F\left(\mathcal{S}_t^C\right) \; if \; \widetilde{E}_t^V\left(\mathcal{W}_{t-1}, \mathcal{S}_{t-1}^C\right) > 0 \tag{5.45}$$

where A_t^T, A_t^F represent the targeted and forwarded values, respectively for the MLPNN action function. It is worth reminding that the MLPNN function for the action space is updated only if the critic error is positive. This leads in fact to a higher learning speed as will be indicated in Chapter 6. The action-value error can be calculated according to Eq. 5.46:

$$\widetilde{E}_t^A\left(\mathcal{W}_{t-1}^A, \mathcal{S}_{t-1}^C, \mathcal{A}_t\right) = \frac{1}{2}\left[A_t^T\left(\mathcal{S}_{t-1}^C\right) - A_t^F\left(\mathcal{S}_{t-1}^C\right)\right]^2 \tag{5.46}$$

The CACLA reasoning is shown by Algorithm 5.5 where the critic decision is pointed by Line 12 and the actor is updated at Line 9.

CACLA can be used for discrete decisions if the output action space is discretized by using the principles from Chapter 4 where the RBFNN function is used in the CQI classification. This approach attracts several drawbacks:

- If the MLPNN falls into the local optima or overestimates the action values, then the entire structure is affected and the scheduling rules are not selected properly. When one MLPNN is used for each discrete action (Q, SARSA, QV, ACLA), if one structure is affected, other functions can be applied by avoiding the usage of the affected MLPNN, which in fact improves the performance of the scheduling procedure.
- The approach presents a limited scalability, and the pool size of scheduling rules has to be decided based on the continuous action space dimension.

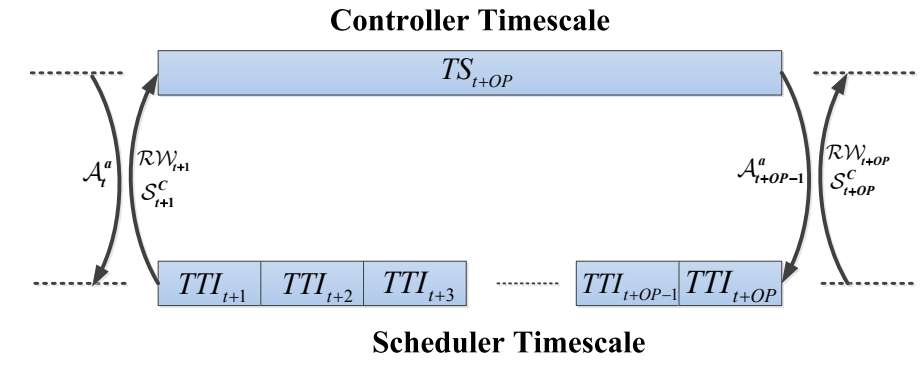
For the DSR-CMOO problems with homogeneous traffic types, it is preferable to use multi-agent systems with specific cooperation between the fairness and QoS agents. The CACLA learning is used as a fairness agent by selecting the continuous action which parameterizes the GPF scheduling rule. For other QoS objectives, the RL algorithms with discrete action spaces can be used.

Algorithm 5.5 CACLA-Learning Based DSR-SMOO MDP

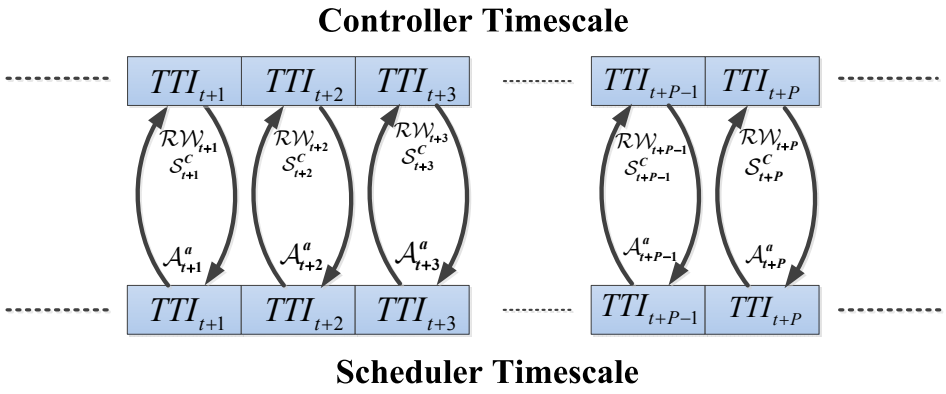
1. **for** each TTI t
2. **Scheduler State Space Aggregation**: $\mathcal{S}_t^S \xrightarrow{Agg.} \mathcal{S}_t^C$
3. **MOO Evaluation procedure**: $\mathcal{RW}_t \leftarrow \mathcal{F}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}\right)$
4. **verify** if $\mathcal{S}_t^C$ is terminal $\left(\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}\right) = 1\right)$
5. **explore** $\mathcal{S}_t^C$ and determine the current continuous action $\mathcal{A}_t : \pi_t^{IE}\left(\mathcal{A}_t \mid \mathcal{S}_t^C\right)$
6. **if** exploitation: $\mathcal{A}_t = \mathcal{F}_{MLP}^A\left(\mathcal{S}_t^C\right)$
7. **else** $\mathcal{A}_t = \arg\max\left(\pi_t^I\left(\mathcal{A}_t \mid \mathcal{S}_t^C\right)\right)$
8. **feed-forward** $V_t^F\left(\mathcal{S}_t^C\right) = \mathcal{F}_{MLP}^V\left(\mathcal{S}_t^C\right)$
9. **update** $V_t^T\left(\mathcal{S}_{t-1}^C\right) \xleftarrow{\eta^V} \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C\right) + \gamma \cdot V_t^F\left(\mathcal{S}_t^C\right)$
10. **back-propagate error :** $\widetilde{E}_t^V\left(\mathcal{W}_{t-1}, \mathcal{S}_{t-1}^C\right)$
11. **update** MLPNN weights**:** $\mathcal{W}_{t,w} = \mathcal{W}_{t-1,w} - \eta_t^V \cdot \partial \widetilde{E}_t^V\left(\mathcal{W}_{t-1,w}, \mathcal{S}_{t-1}^C\right) / \partial \mathcal{W}_{t-1,w}$
12. **if** $\widetilde{E}_t^V\left(\mathcal{W}_{t-1}, \mathcal{S}_{t-1}^C\right) > 0$
13. **update** $A_t^T\left(\mathcal{S}_{t-1}^C\right) \leftarrow A_t^F\left(\mathcal{S}_t^C\right)$
14. **back-propagate error:** $\widetilde{E}_t^A\left(\mathcal{W}_{t-1}^A, \mathcal{S}_{t-1}^C, \mathcal{A}_t\right)$
15. **update** weights**:** $\mathcal{W}_{t,w}^A = \mathcal{W}_{t-1,w}^A - \eta_t^A \cdot \partial \widetilde{E}_t^A\left(\mathcal{W}_{t-1,w}^A, \mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}\right) / \partial \mathcal{W}_{t-1,w}^A$
16. **end if**
17. **MUTI: map** $\mathcal{A}_t$ **to scheduling rule** $c_{o,w_o}[t]$
18. $c_{o,w_o}[t] \rightarrow$ **calculate metrics, RBs allocation and TB calculation**
19. **end for**

5.6.7 The Experience Replay in DSR-SMOO MDP Problems

The behavior of the RRM entities in LTE scheduling based on the RL approach is very important from the viewpoint of the input observations. As mentioned earlier, the controller may decide to spend more time in a given zone of the state space. Since the MLPNN structure is trained over the stochastic gradient descent algorithm, the RRM environment has to provide as many observations as possible from different regions of the aggregate scheduler state space. These aspects can be in contradiction since the quality of the provided



a)



b)

Fig. 5.8 Controller Time Scales a) for the Observation Period of OP > 1TTI and b) for the Observation Period of OP = 1TTI)

observations in the experience replay stage for the MLPNN weight corrections depends on the learned controller policy in the exploration stage.

If the scheduler controller explores more in a particular state space region rather than in other zones, the training data is the subject of under-fitting symptom and get stuck into the local optimum solution. On the other hand, by training more on a certain region of the controller state space, the MLPNN functions for each discrete action have better chances to be explored.

Other concern of the RRM environment refers to the transitions from one particular zone of the state space to another. In other words, when the exploring subspace moves drastically to another one, the controllable parameters can fluctuate and thus, the RRM entity can provide high rewards under the state values which are far away from the optimal state space region. In order to avoid

this drawback, the controller and the scheduler time scales have to be set differently as shown in Fig. 5.8 in order to stabilize the observations from the scheduler state space. When performing different scheduling rules TTI-by-TTI (Fig. 5.8.b), the fluctuations and the error in the reward function can be significant if the MOO evaluation parameters cannot be adapted based on the state condition. Then, the idea of giving the observation period (OP) to the controller becomes mandatory. In other words, the scheduler controller waits until the reward function becomes stable, and then, at the beginning of the new observation period, a new action is applied and the received reward value can be reinforced. Meanwhile, the scheduler uses the same action which was decided in the previous observation time instant. This approach attracts a consistent drawback since the number of RL updates is decreased drastically and the controller needs more time to collect enough observations. Therefore, the solution proposed in Fig. 5.8.a can provide feasible scheduling policies at the price of a longer exploration period.

It is clear that for some policy convergence criteria, the undesired transitions between the state regions cannot be avoided through the exploration process. The proposed method of selecting the input observations for the scheduler controller in order to avoid the under-fitting and the reward fluctuations is based on the following concepts:

- Some parameters can be adjusted in order to make the transition as short as possible. For example, for the DSR-SMOO MDP problems focusing on fairness, if the action steps are small enough, then the transition from one TTI to another cannot be that high.
- ***Experience Replay (ER)***: The scheduler controller can save the convenient tuple $\left(\mathcal{S}_{t-1}^{C}, \mathcal{A}_{t-1}^{a}, \mathcal{RW}_{t}, \mathcal{S}_{t}^{C}, \mathcal{A}_{t}^{a'}\right)$ which does not include the undesired transitions from the previous controller state $\mathcal{S}_{t-1}^{C}$ to the current controller state $\mathcal{S}_{t}^{C}$. An additional processing unit is required in the exploration stage in order to save some observations with good characteristics from different regions of the controller state space. The idea is to store as many observations as possible depending on the memory capacity. At the same time, the MDP observations when the reward is

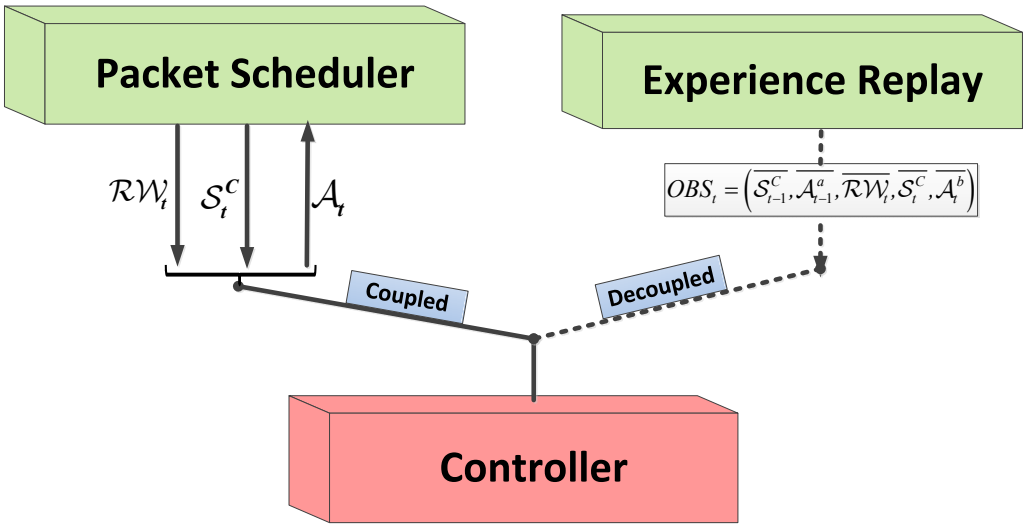


Fig. 5.9 Coupled/Decoupled Interaction Between the Controller and the LTE Scheduler (Experience Replay)

maximized should be saved in order to localize the feasible scheduler states under various system conditions.

The ER stage cannot be mixed with the exploration period and then, it must be performed as an individual stage after the exploration procedure. Based on the ER stage, the interaction between the controller and the scheduler is divided into two directions as shown in Fig. 5.9 and explained bellow:

1. ***Coupled***: This is the case when the controller is trained online based on the observations provided from the LTE scheduler and it corresponds to the exploration and exploitation stages.
2. ***Decoupled***: When the ER stage is performed, the MDP controller is connected to the observation storage entity which provides the visited observations with the desired characteristics.

The observation vector $OBS_t = \left(\overline{\mathcal{S}_{t-1}^C}, \overline{\mathcal{A}_{t-1}^a}, \overline{\mathcal{RW}_t}, \overline{\mathcal{S}_t^C}, \overline{\mathcal{A}_t^b}\right)$ is randomly selected at each TTI from the database and $\overline{\mathcal{S}_{t-1}^C}$ is the stored element during the exploration stage. The stored set should contain all observations when the scheduler reward is $\mathcal{RW} = 1$ in order to keep the correct direction to the optimal policies. Different methods of selecting the observations for the ER stage can be implemented. One example of such approaches is to pick up, at every 10 updates, one update that

Algorithm 5.6 ER-QV-Learning Based DSR-SMOO MDP

if $Decoupled() = true$
  for each TTI t
    percept : $OBS_t = \left(\overline{\mathcal{S}_{t-1}^C}, \overline{\mathcal{A}_{t-1}^a}, \overline{\mathcal{RW}_t}, \overline{\mathcal{S}_t^C}, \overline{\mathcal{A}_t^b}\right)$
    feed-forward $V_t^F\left(\overline{\mathcal{S}_t^C}\right) = \mathcal{F}_{MLP}\left(\overline{\mathcal{S}_t^C}\right)$
    update $V_t^T\left(\overline{\mathcal{S}_{t-1}^C}\right) \xleftarrow{\eta^V} \overline{\mathcal{RW}_t}\left(\overline{\mathcal{S}_{t-1}^C}\right) + \gamma \cdot V_t^F\left(\overline{\mathcal{S}_t^C}\right)$
    back-Propagate Error $\widetilde{E}_t^V\left(\mathcal{W}_{t-1}^a, \overline{\mathcal{S}_{t-1}^C}\right)$
    update MLPNN weights**:** $\mathcal{W}_{t,w} = \mathcal{W}_{t-1,w} - \eta_t^V \cdot \partial \widetilde{E}_t^V\left(\mathcal{W}_{t-1,w}, \overline{\mathcal{S}_{t-1}^C}\right) / \partial \mathcal{W}_{t-1,w}$
    update $Q_t^T\left(\overline{\mathcal{S}_{t-1}^C}, \overline{\mathcal{A}_{t-1}^a}\right) \xleftarrow{\eta^Q} \overline{\mathcal{RW}_t}\left(\overline{\mathcal{S}_{t-1}^C}, \overline{\mathcal{A}_{t-1}^a}\right) + \gamma \cdot V_t^F\left(\overline{\mathcal{S}_t^C}\right)$
    back-propagate error: $\widetilde{E}_t^Q\left(\mathcal{W}_{t-1}^a, \overline{\mathcal{S}_{t-1}^C}, \overline{\mathcal{A}_{t-1}^a}\right)$
    update weights**:** $\mathcal{W}_{t,w}^a = \mathcal{W}_{t-1,w}^a - \eta_t^Q \cdot \partial \widetilde{E}_t^Q\left(\mathcal{W}_{t-1,w}^a, \overline{\mathcal{S}_{t-1}^C}, \mathcal{A}_{t-1}^a\right) / \partial \mathcal{W}_{t-1,w}^a$
  end TTI t
end if

represents the optimal state space. Unfortunately, there is not any specific rule, and the best way is to use the greedy policy in the selection of the saved observations. The RL algorithms keep the same reasoning as indicated in previous section. Algorithm 5.6 indicates the ER based QV-learning principle. The same concepts can be applied for other RL algorithms except CACLA.

5.7 The Reinforcement Learning for DSR-CMOO Focusing on Fairness and QoS Objectives

The RL algorithms discussed in the previous sections address the problem of scheduling decisions with the homogenous traffic type focusing on one objective. The purpose of this section is to propose a feasible model for multiple scheduling objectives when the RL approach is applied. If the traffic type is still homogenous and the scheduling rules which are focused on different QoS targets belong to the same discrete action set $\mathcal{A}_t$, then the RL principles remain unchanged when compared with the DSR-SMOO case. The MDP definition keeps

the same form of $(MDP):\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{a},\mathcal{RW}_{t}^{MO},\mathcal{S}_{t}^{C},\mathcal{A}_{t}^{b}\right)$ with a difference that the multi-objective reward function computation implies:

$$\mathcal{RW}_{t}^{MO}=\delta_{T}\cdot\mathcal{RW}_{t}^{T}+\delta_{F}\cdot\mathcal{RW}_{t}^{F}+\delta_{G}\cdot\mathcal{RW}_{t}^{G}+\delta_{D}\cdot\mathcal{RW}_{t}^{D}+\delta_{P}\cdot\mathcal{RW}_{t}^{P}+\delta_{S}\cdot\mathcal{RW}_{t}^{S} \tag{5.47}$$

where $\mathcal{RW}_{t}^{MO}$ represents the total reward, $\mathcal{RW}_{t}^{T}$ is the system throughput reward, $\mathcal{RW}_{t}^{F}$ is the fairness reward, $\mathcal{RW}_{t}^{G}$ is the reward which indicates the GBR constraint performance, $\mathcal{RW}_{t}^{D}$ is the reward for the HoL objective, $\mathcal{RW}_{t}^{P}$ represents the reward for the PLR (or PDR) performance, and finally, $\mathcal{RW}_{t}^{S}$ indicates if the scheduler is stable in terms of the active queues. The set of parameters $\left\{\delta_{T},\delta_{F},\delta_{G},\delta_{D},\delta_{P},\delta_{S}\right\}\in\left[0,1\right]$ permits to set the importance of each particular objective in the total reward. If all the particular sub-rewards have the same characteristics and can merge to the optimal value of one, then the overall DSR-CMOO MDP problem can be considered to be episodic.

As mentioned, the proposed central controller is able not only to select the scheduling rule but also to parameterize some marginal functions. For instance, the GPF is responsible for the fairness performance and $\mathcal{RW}_{t}^{F}$ is issued based on the parameterization performance of GPF in the previous TTI. When a marginal function needs to be parameterized, the action set $\mathcal{A}_{t}$ has to be divided into two subsets: one discrete action subset represents the marginal function index and the other subset represents the necessary steps involved in the GPF parameterization. If the second action space is decided to be continuous and similar to the CACLA case, two agents have to work in the specific cooperation manner as follows:

- **<u>QoS Agent:</u>** defined by the MDP problem $\left(\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{Q,a},\mathcal{RW}_{t}^{MO},\mathcal{S}_{t}^{C},\mathcal{A}_{t}^{Q,b}\right)$ where $\mathcal{A}_{t}^{Q,b}$ is the discrete action addressing the scheduling rule focusing on QoS objectives such as fairness, stability, HoL delay, GBR or PDR.
- **<u>Fairness Agent</u>**: defined by the MDP problem $\left(\mathcal{S}_{t-1}^{C,F},\mathcal{A}_{t-1}^{F},\mathcal{RW}_{t}^{F},\mathcal{S}_{t}^{C,F},\mathcal{A}_{t}^{F}\right)$ where $\mathcal{A}_{t}^{F}$ is the continuous action space corresponding to the CACLA TD learning approach. The fairness state space is extracted from the

overall controller state space $\mathcal{S}_{t-1}^{C}$. More details about these novel concepts are provided in Chapter 7.

Based on the fairness parameterization of each scheduling rule analyzed in Chapter 3, the fairness and QoS agents can work under two main modes:

1. **GPF-SP/DP Fairness Parameterization**: In this case, the fairness parameter set (α_t, β_t) is adapted only for the GPF-SP/DP scheduling rules, whereas other QoS based rules maintain a constant level of these parameters $(\alpha = 1, \beta = 1)$. The fairness agent is selected only when the QoS agent decides to improve the fairness performance (with specific cooperation). Then, the QoS agent reinforces the reward value of $\mathcal{RW}_t^{MO}$ whereas the fairness agent reinforces the fairness reward $\mathcal{RW}_t^{F}$. When the GPF-SP/DP rule is not selected, then the fairness agent is on the idle phase. Details about this innovative concept are provided in Chapter 7.
2. **QoS Objectives Based GPF-SP/DP Fairness Parameterization**: (e.g. GPF-RAD or GPF-MLWDF with (α_t, β_t) adaptation): In this case, all the scheduling rules adapt the fairness parameters. The fairness and QoS agents can perform in parallel without cooperation by using different state spaces, action spaces and reward functions. The QoS agent is rewarded with the total reward $\mathcal{RW}_t^{MO}$ and the fairness agent reinforces only the fairness reward $\mathcal{RW}_t^{F}$. In this case, the fairness agent is active at each TTI.

Figure 5.10 highlights the basic architecture of the DSR-CMOO MDP problems with double agents by following the first mode being exposed above with dynamic fairness parameterization for the GPF-SP/DP rules and static fairness parameterization for other QoS based GPF rules. Different sets of MLPNN functions are used for each agent in order to evaluate the state, action and state-action values for different RL approaches. As seen, the central controller assures the specific cooperation between fairness and QoS agents. The fairness and QoS actions are sent to the MUTI entity which verifies whether or not the scheduling rule is GPF-SP/DP. When the GPF-SP/DP rules are selected, the

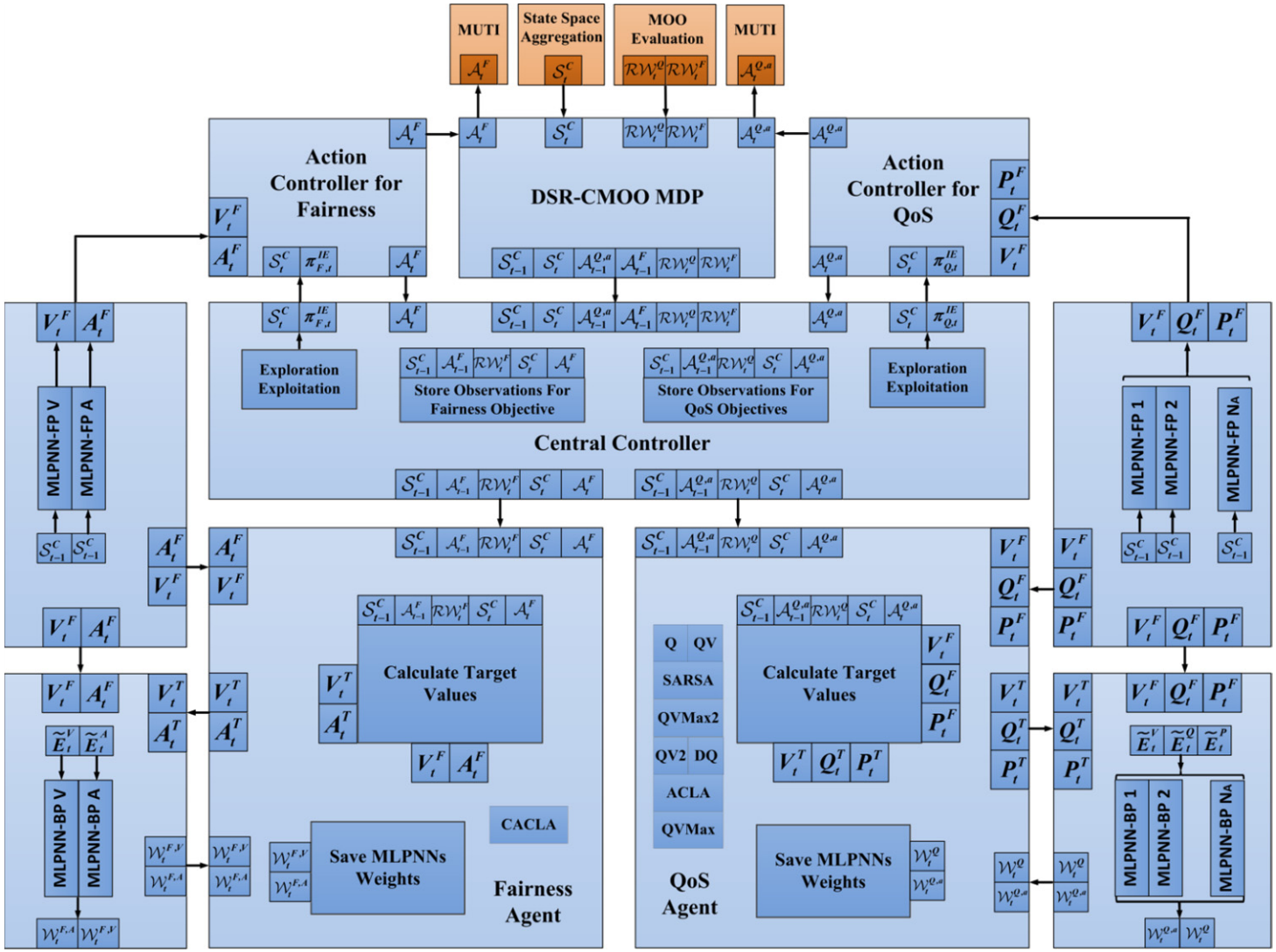


Fig. 5.10 RL Based DSR-CMOO MDP

Algorithm 5.7 ACLA-Learning and CACLA-Learning Based DSR-CMOO MDP

1. for each TTI t

2. Scheduler State Space Aggregation: $\mathcal{S}_t^S \xrightarrow{Agg.} \mathcal{S}_t^C$

3. MOO Evaluation procedure: $\mathcal{RW}_t^{MO} \leftarrow \mathcal{F}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^{Q,a}, \mathcal{A}_{t-1}^F\right)$

4. **verify** if $\mathcal{S}_t^C$ is terminal $\left(\mathcal{RW}_t^{MO}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^{Q,a}, \mathcal{A}_{t-1}^F\right)=1\right)$

5. **explore** $\mathcal{S}_t^C$ for all actions $\mathcal{A}_t^{Q,a'}$, $a'=1,..,|\mathcal{A}|$ based on $\pi_t^{IE}\left(\mathcal{A}_t^{Q,a'} \mid \mathcal{S}_t^C\right)$

6. **if** is policy evaluation (MLPNN FP): $\mathcal{A}_t^{Q,b} = \arg\max_{a'}\left(\mathcal{F}_{MLP}^{Q,a'}\left(\mathcal{S}_t^C, \mathcal{A}_t^{Q,a'}\right)\right)$

7. **else** is policy improvement: $\mathcal{A}_t^{Q,b} = \arg\max_{a'}\left(\pi_t^I\left(\mathcal{A}_t^{Q,a'} \mid \mathcal{S}_t^C\right)\right)$

8. **if** $\left(\mathcal{RW}_t^{MO}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^{Q,a}\right)=1\right)$ store observations $\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^{Q,a}, \mathcal{RW}_t^{MO}, \mathcal{S}_t^C, \mathcal{A}_t^{Q,b}\right)$

9. **else if** $(1-\varepsilon) < \varepsilon_{INIT}$ store observations $\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^{Q,a}, \mathcal{RW}_t^{MO}, \mathcal{S}_t^C, \mathcal{A}_t^{Q,b}\right)$

10. **if** $\left(\mathcal{A}_{t-1}^{Q,a} \rightarrow GPF - SP / DP\right)$

11. **verify** if $\mathcal{S}_t^{C,F}$ is terminal for Fairness Objective $\left(\mathcal{RW}_t^F\left(\mathcal{S}_{t-1}^{C,F}, \mathcal{A}_{t-1}^F\right)=1\right)$

12. **explore** $\mathcal{S}_t^{C,F}$ for $\mathcal{A}_t^F : \pi_t^{IE}\left(\mathcal{A}_t^F \mid \mathcal{S}_t^{C,F}\right)$

13. **if** exploitation: $\mathcal{A}_t^F = \mathcal{F}_{MLP}^{F,A}\left(\mathcal{S}_t^{C,F}\right)$

14. **else** $\mathcal{A}_t^F = arg\,max\left(\pi_t^I\left(\mathcal{A}_t^F \mid \mathcal{S}_t^{C,F}\right)\right)$

15. **feed-forward** $V_t^{F,F}\left(\mathcal{S}_t^{C,F}\right) = \mathcal{F}_{MLP}^{F,V}\left(\mathcal{S}_t^{C,F}\right)$

16. **update** $V_t^{F,T}\left(\mathcal{S}_{t-1}^{C,F}\right) \xleftarrow{\eta^V} \mathcal{RW}_t^F\left(\mathcal{S}_{t-1}^{C,F}\right) + \gamma \cdot V_t^{F,F}\left(\mathcal{S}_t^{C,F}\right)$

17. **back Propagate Error :** $\widetilde{E}_t^{F,V}\left(\mathcal{W}_{t-1}^{F,V}, \mathcal{S}_{t-1}^{C,F}\right)$

18. **update** weights: $\mathcal{W}_{t,w}^{F,V} = \mathcal{W}_{t-1,w}^{F,V} - \eta_t^V \cdot \partial \widetilde{E}_t^{F,V}\left(\mathcal{W}_{t-1,w}^{F,V}, \mathcal{S}_{t-1}^{C,F}\right) / \partial \mathcal{W}_{t-1,w}^{F,V}$

19. **if** $\widetilde{E}_t^{F,V}\left(\mathcal{W}_{t-1}^{F,V}, \mathcal{S}_{t-1}^{C,F}\right) > 0$

20. **update** $A_t^{F,T}\left(\mathcal{S}_{t-1}^{C,F}\right) \leftarrow A_t^{F,F}\left(\mathcal{S}_t^{C,F}\right)$

21. **back Propagate Error:** $\widetilde{E}_t^{F,A}\left(\mathcal{W}_{t-1}^{F,A}, \mathcal{S}_{t-1}^{C,F}, \mathcal{A}_t^F\right)$

22. **end if**

23. **update** weights: $\mathcal{W}_{t,w}^{F,A} = \mathcal{W}_{t-1,w}^{F,A} - \eta_t^A \cdot \partial \widetilde{E}_t^{F,A}\left(\mathcal{W}_{t-1,w}^{F,A}, \mathcal{S}_{t-1}^{C,F}, \mathcal{A}_{t-1}^F\right) / \partial \mathcal{W}_{t-1,w}^{F,A}$

24. **end if**

25. **feed-forward** $V_t^{Q,F}\left(\mathcal{S}_t^C\right) = \mathcal{F}_{MLP}^{Q}\left(\mathcal{S}_t^C\right)$

26. **update** $V_t^{Q,T}\left(\mathcal{S}_{t-1}^C\right) \xleftarrow{\eta^V} \mathcal{RW}_t^T\left(\mathcal{S}_{t-1}^C\right) + \gamma \cdot V_t^{Q,F}\left(\mathcal{S}_t^C\right)$

27. **back-Propagate Error :** $\widetilde{E}_t^{Q,V}\left(\mathcal{W}_{t-1}^{Q,V}, \mathcal{S}_{t-1}^C\right)$

28. **update** weights: $\mathcal{W}_{t,w}^{Q,V} = \mathcal{W}_{t-1,w}^{Q,V} - \eta_t^V \cdot \partial \widetilde{E}_t^{Q,V}\left(\mathcal{W}_{t-1,w}^{Q,V}, \mathcal{S}_{t-1}^C\right) / \partial \mathcal{W}_{t-1,w}^{Q,V}$

29. **Determine the critic decision** $\mathcal{P}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^{Q,a}\right)$ **based on Eq. 5.44**

30. **update** $P_t^{Q,T}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^{Q,a}\right) \xleftarrow{\eta^Q} \mathcal{P}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right) - P_t^{Q,F}\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$

31. **back Propagate Error:** $\widetilde{E}_t^{Q,P}\left(\mathcal{W}_{t-1}^{Q,a},\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{Q,a}\right)$

32. **update** weights: $\mathcal{W}_{t,w}^{Q,a}=\mathcal{W}_{t-1,w}^{Q,a}-\eta_t^P\cdot\partial\widetilde{E}_t^{Q,P}\left(\mathcal{W}_{t-1,w}^{Q,a},\mathcal{S}_{t-1}^{C},\mathcal{A}_{t-1}^{Q,a}\right)/\partial\mathcal{W}_{t-1,w}^{Q,a}$

33. **MUTI: map** $\left(\mathcal{A}_t^{Q,b},\mathcal{A}_t^{F}\right)$ **to scheduling rule** $c_{o,w_o}[t]$

34. $c_{o,w_o}[t]\rightarrow$ **calculate metrics, RB allocation and TBS calculation**

35. **end** TTI t

continuous action space $\mathcal{A}_t^F$ is considered. If the GPF scheduling rule with simple parameterization is taken into account, then the action space dimension for CACLA learning is $\left|\mathcal{A}_t^F\right|=1$. When the GPF rule with double parameterization is applied, the action space dimension is $\left|\mathcal{A}_t^F\right|=2$. For the QoS objectives, the RL algorithms with discrete actions can be used. Algorithm 5.7 explains the exploration principles of using the CACLA learning for the fairness agent and ACLA algorithm for the QoS agent. The proposed algorithm follows the principles of CACLA and ACLA for the DSR-SMOO MDP problems presented in the previous sub-sections. The details about the proposed DSR-CMOO MDP architecture for the heterogeneous traffic types are analyzed in Chapter 8.

5.8 Summary

In this chapter, the principles of RL algorithms in LTE packet scheduling have been discussed. At the very basic level, the DSR-SMOO/CMOO scheduling can be seen as a MDP problem in which the scheduling decision in the current TTI depends only on the decision applied in the previous time instant. The contribution of this chapter refers to the possibility of introducing more complex and realistic packet scheduling problems which are focused on the multi-objective criteria. The DSR-CMOO optimization problems involve in fact the MARL architecture with a specific cooperation between two agents: fairness and QoS. The QoS agent selects different scheduling rules being oriented on different objectives. When the GPF-SP/DP rules corresponding to the fairness performance are selected, then the fairness agent explores the fairness sub-space and reinforces its corresponding reward. The advantage of the proposed architecture refers to the

fact that the fairness policy can be learned first in order to provide enough observations for the optimal fairness decisions. Afterwards, the QoS policy can be learned by exploiting, when necessary, the learned fairness scheduling policy.

The principles highlighted in this chapter are used in Chapter 6 and Chapter 7 to optimize the sustainable scheduling policies focusing on particular and multiple scheduling objectives.

Chapter 6 solves the DSR-SMOO MDP problems for the NGMN fairness criterion and for GBR requirements by using the RL algorithms from Sub-section 5.6.1 to Sub-section 5.6.6. When the NGMN fairness criterion is considered, CACLA2 which parameterizes the GPF-DP scheduling rule is able to outperform the entire set of scheduling policies obtained by using other RL approaches. When the CACLA2 scheduling policies are exploited, the number of feasible TTIs increases to more than 95% from the total number of simulated TTIs, denoting in this sense the sustainability of the proposed policies. When the GBR requirements are considered, the RL algorithms with discrete action space are used in order to select the most suitable scheduling rule focusing on GBR objective at each TTI. In this case, ACLA policy provides much better performance when compared with the static scheduling rules by maximizing at the same time, the number of TTIs when all active users are 100% satisfied from the viewpoint of the GBR objective.

Chapter 7 proposes sets of sustainable scheduling policies which solve the DSR-CMOO problems being focused on NGMN fairness requirement, GBR, PDR and HoL packet delay objectives. The architecture from Fig. 5.10 is exploited in order to find the best scheduling rule which is able to maximize the multi-objective reward at each TTI. CACLA2+ is a novel approach being able to parameterize the GPF-DP scheduling rule and to adjust the filter length which is used in the AUT-MMF computations. The scheduling policy obtained by combining SARSA (QoS policy) and CACLA2+ (fairness policy) is able to outperform other methodologies when the multi-objective target is considered, by maximizing at the same time, the number of feasible TTIs when the VBR and CBR traffic types are simulated.

Chapter 6

Sustainable Scheduling Policies for Sequential Multi-Objective Optimization

6.1 Chapter Outline

The DSR-SMOO MDP problems are studied in this chapter in terms of NGMN fairness criterion and GBR constraint objective. For both objectives, there are two methods of computing the AUT observations for the RL reward functions: exponential or median moving filters. When the exponential filter is used, the advantage of using the CQI aggregation techniques from Chapter 4 is studied for the NGMN objective. It is shown that the channel information is undoubtedly necessary in order to outperform other existing technologies by maximizing the percentage of TTIs when the scheduler is feasible and by minimizing, at the same time, the amount of TTIs when the scheduler stays in one of the undesirable states such as *unfair* or *over-fair*. When the median moving filter is used, the optimality of different filter lengths is studied by training different MLPNN functions with various RL approaches. The windowing factor is used in this sense to set different filter time window lengths. By using extensive simulation results, the optimum windowing factor domain is proposed in order to maintain the integrity of the learned policies. It is proven that the exploited scheduling policies which solve the simple or double GPF parameterization problem by using CACLA1/CACLA2 RL algorithms are able to outperform the existing methodologies from the viewpoint

of the NGMN fairness constraint for restrictive, optimum or large windowing factor domains. The DSR-SMOO problem being focused on GBR constraint mixes at each TTI four scheduling rules: three existing scheduling disciplines and the novel scheduling rule which updates the Lagrange multiplier for user throughputs. Alongside the windowing factor settings, the type of used traffic plays a crucial role in satisfying all active bearers from the GBR constraint point of view. The optimum windowing factor should then be found for each considered traffic model in order to keep the integrity of the learning procedure. When the learned policies which solve the DSR-SMOO MDP problems being focused on the GBR objective are applied, it is shown that ACLA actor-critic learning schemes increase the amount of TTIs if all active bearers are 100% GBR satisfied when compared with the classical scheduling rules and minimize at the same time, the number of punishment rewards for the full buffer, CBR and VBR traffic types.

6.2 DSR-SMOO MDP Focusing on NGMN Fairness Objective

As mentioned in Chapter 3, the target of the proposed scheduling approach is to maximize the aggregate objective function concurrently on the short term purpose by selecting the scheduling rule with the highest reward value at each TTI. The user fairness performance is strongly connected with other objectives such as GBR, HoL packet delay, PLR constraints and queue stability criterion. When the optimal controller state is reached in terms of the aforementioned objectives, the level of fairness of active bearers becomes even more important. Without affecting the optimality of other objectives, the user fairness performance can be improved in order to provide radio resources to the pending radio bearers. In this sense, a proper tradeoff between user fairness and system throughput should be found in order to maintain the optimality of other objectives and to maximize the scheduler capacity of accepting new bearers. Therefore, the particular objective of user fairness should be addressed first in order to propose a novel method for the tradeoff adaptation. The simple and double GPF

parameterization problems focusing on the NGMN fairness requirement are studied in this section by using a set of scheduling policies obtained when performing different RL algorithms in the exploration stage.

6.2.1 Tradeoff Between System Throughput and User Fairness

The tradeoff level between system throughput and user fairness can be measured by using proper fairness performance metrics and achieved by using optimal parameterization of the GPF scheduling rule. In the current proposal, the fairness performance metric provides the reward value of the RL algorithms and the proper GPF parameter (or parameters) is (or are) mapped from the optimal controller actions. Based on the principles shown in Chapter 5, the aim of this study is to use the MLPNN function approximation in order to generalize a non-linear function which practically maps the aggregate controller state space into a desired and optimal GPF decision. The non-linear fairness function is learned based on the RL algorithms discussed in Chapter 5 by reinforcing the MLPNN output error between the target state-action, action and state values and the desired output values. Therefore, the aggregate controller state space, the action set and the reward function should be defined in order to learn the optimal policies which focus on the user fairness requirement.

The optimal radio resource allocation is assured by the optimization problem analyzed in Eq. 3.35 in Chapter 3, where the maximization operator respects the Shannon capacity limit under certain fairness performance requirement. Based on the CQI state space module, the set of achievable user rates is obtained by using the mapping procedure from the SINR levels to the spectral efficiency values highlighted in Table B.1 from Appendix B. When a given scheduling rule is applied, the obtained sum of instantaneous rates is entitled Instantaneous System Throughput (IST). As mentioned, the IST is optimal at each TTI based on the fairness optimization problem which aims to maximize the sum of GPF metrics for certain number of resource blocks.

One performance fairness criterion claims to allocate in each TTI equal number of RBs to each active bearer without taking the channel conditions into account. In this case, the IST is decreased when compared with the optimal case and it represents the sum of equally distributed instantaneous rates for each user. The scheduling rule is entitled Round Robin (RR) and it is considered to be sub-optimal from the viewpoint of RB allocation. Therefore, the tradeoff level between IST and user fairness cannot be addressed since the system throughput is seriously degraded over the time. In order to avoid this aspect, the system throughput should be averaged over a given time window, known as Average System Throughput (AST). The AST value is obtained at each TTI by averaging the instantaneous user rates, measure which is entitled Average User Throughput (AUT). By averaging the IST value, the fairness criteria can be achieved in a given number of TTIs. This flexibility involves in fact the optimal resource allocation based on GPF rule which practically aims to allocate resources fairly over the time, whereas the RR scheduling rule allocates the resources instantaneously in the frequency domain. The fairness flexibility involves two ways in evaluating the tradeoff between AST and fairness, such as:

- **Short Term Fairness (STF)**: The time window used in averaging the instantaneous user rate is reduced (e.g. 10 TTIs). Based on this performance criterion, the tradeoff value is immediately affected if the conditions (e.g. channel conditions, number of active radio bearers) change dramatically from one TTI to another. Under very restrictive fairness constraints, AST is expected to be decreased, but the allocation of RBs will stay optimal if the GPF rule is performed.
- **Long Term Fairness (LTF)**: The time window is large enough to permit in fact higher system throughput with the respect of fairness constraint on a longer term purpose. However, the scheduler controller may not be able to control the tradeoff performance when a very large window is used since the reward value may contain the contribution of several actions being applied on the scheduler states for the considered time window.

The balance between STF and LTF strongly depends on the CMOO performance at each TTI. If the optimal controller state is reached in terms of the QoS

objectives, then the AUT time window can be decreased and the fairness level can be increased in order to accept more pending radio bearers. These novel concepts are discussed in detail in Chapter 7.

6.2.2 User Fairness Performance Measures

The averaging procedure of the instantaneous user rates can be achieved with different averaging filters. In Eq. 3.18, the AUT is calculated by using the exponential moving filter (EMF) where $1/\beta_{\overline{T}}$ represents the time window length. If the average user throughput with exponential moving filter (AUT-EMF) increases exponentially, the respective bearer has lesser chances to be scheduled when the GPF-SP/DP rule is performed. In this case, AUT-EMF can be used for the GPF computation as well for the user fairness performance measure. When the STF performance is analyzed based on AUT-EMF, the $\beta_{\overline{T}}$ parameter must be increased and consequently high oscillations are introduced when the scheduling performance is analyzed. This aspect is undesirable since it affects the learning procedure and consequently the policy refinement of scheduling rules.

Another way of obtaining the AUT observations is to store the IUTs for a given time window and to average these values at each TTI by using the Median Moving Filter (MMF). The AUT-MMF computation is expressed by Eq. 6.1:

$$\overline{\overline{T}}_i[t] = \frac{1}{T_w^M} \sum_{x=0}^{T_w^M} T_i[t-x] \tag{6.1}$$

where T_w^M is the median filter length. When T_w^M is very large, the LTF approach is considered whereas when T_w^M is small enough, the short term fairness performance measure is considered. The AUT-MMF can be used as an evaluation metric for the fairness and the satisfaction of GBR objectives.

Based on the target type, the fairness performance measures can be divided in two categories:

1. **Quantitative Measures**: the performance target represents a predefined constraint and the fairness can be improved by allocating more resources

to users with lower AUT-EMF or AUT-MMF degrading at the same time the overall system throughput.

2. **Qualitative Measures:** the fairness requirement is defined in terms of the overall distribution of Normalized AUT-EMF (NAUT-EMF) or Normalized AUT-MMF (NAUT-MMF). Based on the current channel conditions, the scheduling procedure should be conducted in such a way that the distribution of the achieved normalized throughputs lies in a given region of the fairness performance target, in the distribution domain.

For simplicity, let us define the two dimension set such as $\mathcal{T}_i = \left\{ \overline{T}_i, \overline{\overline{T}}_i \right\}$. The normalized set $\widehat{\mathcal{T}}_i$ is determined based on Eq. 4.5 from Chapter 4 for both types of averaged user throughputs.

One of the most well-known quantitative fairness metric being initially used in the shared computer systems is the Jain Fairness Index (JFI) [241]. The JFI performance index is applied in the LTE scheduling procedure by using the AUT-EMF or AUT-MMF observations. Mathematically, the JFI index can be expressed such that [241]:

$$JFI[t] = \frac{\left(\sum_{i=1}^{|\mathcal{U}_t|} \mathcal{T}_i[t] \right)^2}{|\mathcal{U}_t| \cdot \sum_{i=1}^{|\mathcal{U}_t|} \left(\mathcal{T}_i[t] \right)^2} \tag{6.2}$$

The main problem of introducing the $JFI[t]$ metric in the SMOO problems refers to the difficulty of setting a predefined target value. The JFI constraint should be defined and adapted at the same time based on the performance of other objectives. In this sense, the learned policy of GPF parameters is concentrated in achieving the imposed JFI constraints at each TTI without any consideration of the overall distribution of $\widehat{\mathcal{T}}_i$ observations. This aspect leads in fact that some users may be in outage and that they are restricted in receiving resources for a longer time in the detriment of other users with much better channel conditions. The NGMN qualitative fairness measure considers the resource allocation of one active bearer depending on other achieved normalized user throughputs [13]. The

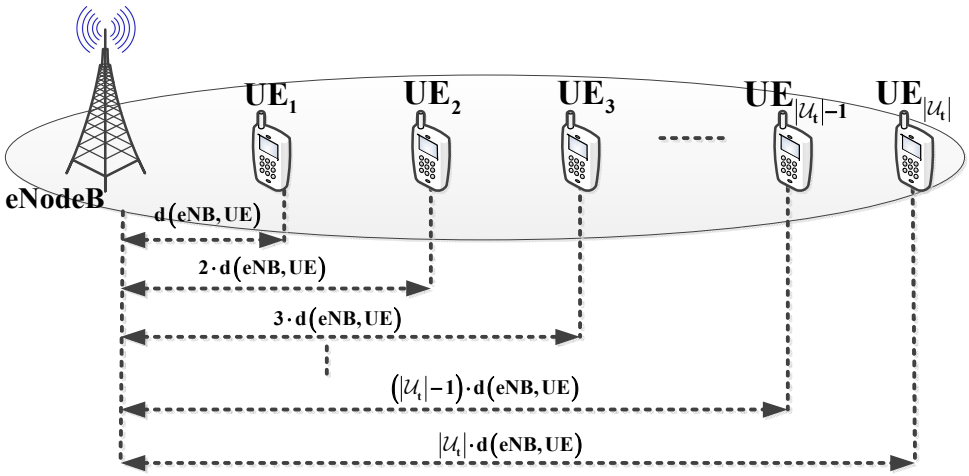


Fig. 6.1 Uniform Physical Distribution (in meters) of 60 Users Scenario Scheduled by Using the GPF Rule with Simple Parameterization

proposed measure aims to evaluate the normalized throughput $\widehat{\mathcal{T}}_i$ distribution TTI-by-TTI and to match the distribution against the NGMN requirement. In this sense, the normalized throughput $\widehat{\mathcal{T}}_i$ bound is defined in the CDF domain, where $\widehat{\mathcal{T}}_i$ should be at least on the right side of this limit. The NGMN fairness condition which should be satisfied at each TTI is highlighted by Eq. 6.3:

$$\psi_i\left(\widehat{\mathcal{T}}_i[t]\right) \le \widehat{\mathcal{T}}_i[t] \tag{6.3}$$

where the cumulative distribution function is calculated based on the log-normal distribution such as [13]:

$$\psi_i\left(\widehat{\mathcal{T}}_i[t]\right) = 0.5 + 0.5 \cdot erf\left[\frac{\ln\left(\widehat{\mathcal{T}}_i[t]\right) - \mu_{\mathcal{T}}}{\sqrt{2}\cdot\sigma_{\mathcal{T}}}\right] \tag{6.4}$$

where the mean $\mu_{\mathcal{T}}$ and STD $\sigma_{\mathcal{T}}$ values are calculated based on Equations 4.6 and 4.7 from Chapter 4. The numerical interpretation of Eq. 6.3 denotes the fact that the normalized throughput of at least $p[\%] = \widehat{\mathcal{T}}_i[t]\%$ must be achieved by at least $(1-p)\cdot|\mathcal{U}_t|\%$ number of users. The adopted fairness measure is inherited from other cellular and wireless standards such as CDMA2000 [242], IEEE 802.16j [243] and IEEE 802.20 [244]. For a more concise explanation, the generic

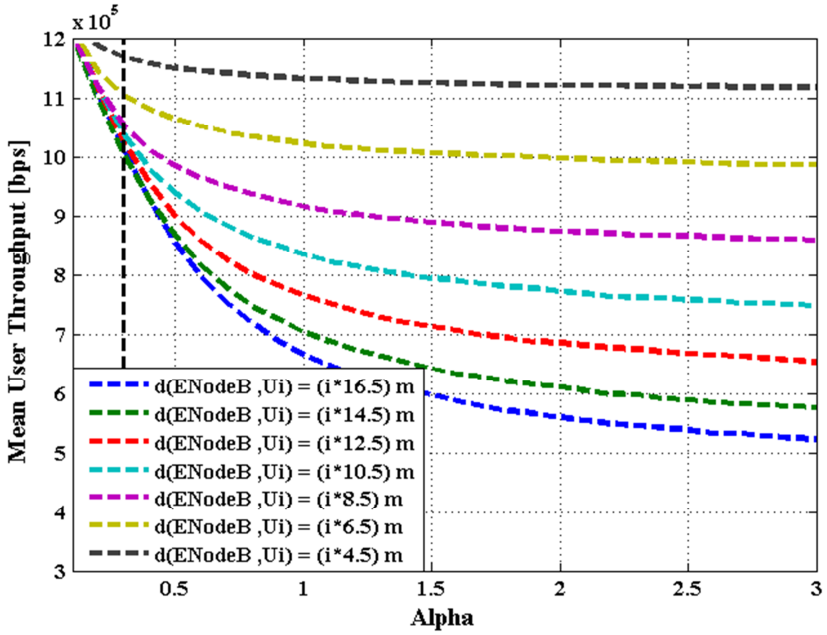


Fig. 6.2.a Mean AST-EMF with $(\alpha - \text{var}, \beta = 1)$ GPF Parameterization $(\beta_{\bar{T}} = 0.01)$ for 60 users scenario being equally distributed from the eNodeB base station to the edge of cell under uniform power allocation and FDD downlink transmission with a system bandwidth of 20MHz

scenario from Fig. 6.1 is analyzed. The AST-EMF, JFI-EMF and $\psi_i\left(\hat{\bar{T}}_i[t]\right)$ performance are discussed for seven scenarios where the physical distances $d(eNB, UE_i)$ are distributed in such a manner that comprises the most relevant general channel conditions. For instance, when $d(eNB, UE_i) = 4.5$, all users are grouped near eNodeB by experiencing good channel conditions whereas the distance of $d(eNB, UE_i) = 16.5$ involves a larger spreading factor of user positions among the cell coverage. The simulation results are conducted through Jakes fading model with a macro-cell urban area on 20MHz system bandwidth with static user position and infinite buffer traffic type. The GPF scheduling rule with simple parameterization $(\alpha - \text{var}, \beta = 1)$ is performed and the main interest is captured by the JFI performance metric in the presence of different α parameters.

From the system throughput point of view (Fig. 6.2.a), it can be seen that the mean AUT-EMF decreases when α increases for the considered scenarios.

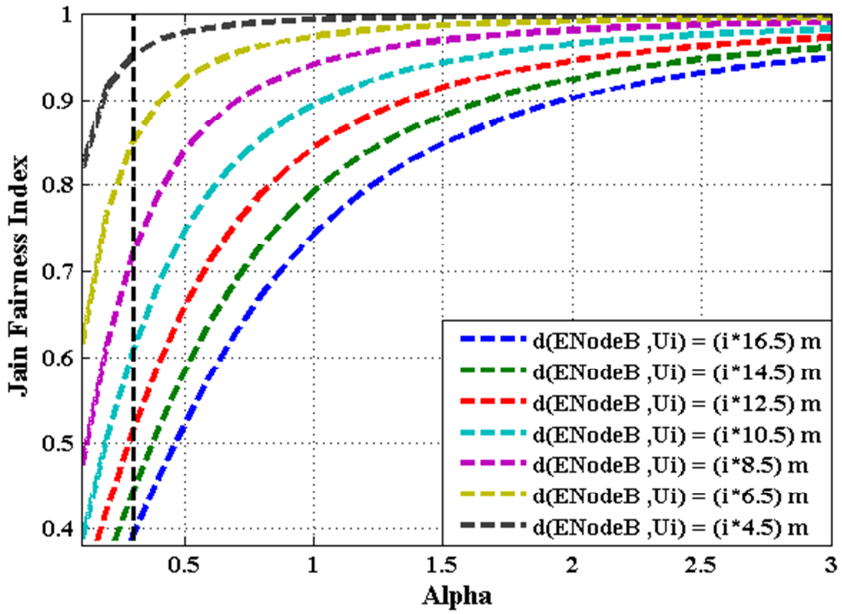


Fig. 6.2.b JFI-EMF with $(\alpha - \text{var}, \beta = 1)$ GPF Parameterization $(\beta_{\bar{T}} = 0.01)$ for 60 users scenario being equally distributed from the eNodeB base station to the edge of cell under uniform power allocation and FDD downlink transmission with a system bandwidth of 20MHz

The system throughput is less deprecated when $\alpha \nearrow$ for the case when users experience very good channel conditions $\left(d\left(eNB, UE_i\right) = 4.5\right)$. The importance of α parameter becomes more visible when more realistic scenarios are considered $\left(d\left(eNB, UE_i\right) = 16.5\right)$. In this case, the overall throughput loss is of about 6Mbps when compared with the first analyzed scenario.

In the case of JFI-EMF performance metric (Fig. 6.2.b), any increase of α parameter leads to higher JFI index performances by degrading the system capacity at the same time. When $\alpha > 1$, the JFI-EMF performance remains relatively constant for the particular case of $d\left(eNB, UE_i\right) = 4.5$.

In Figures 6.2.a and 6.2.b, the reference value of $\alpha = 0.3$ is considered in order to be analyzed for the CDF function $\psi_i\left(\hat{\bar{T}}_i[t]\right)$ performance when the EMF forgetting factor is $\beta_{\bar{T}} = 0.01$. Figure 6.3 plots the CDF variation TTI-by-TTI when the normalized NAUT-EMF distribution is considered. The continuous

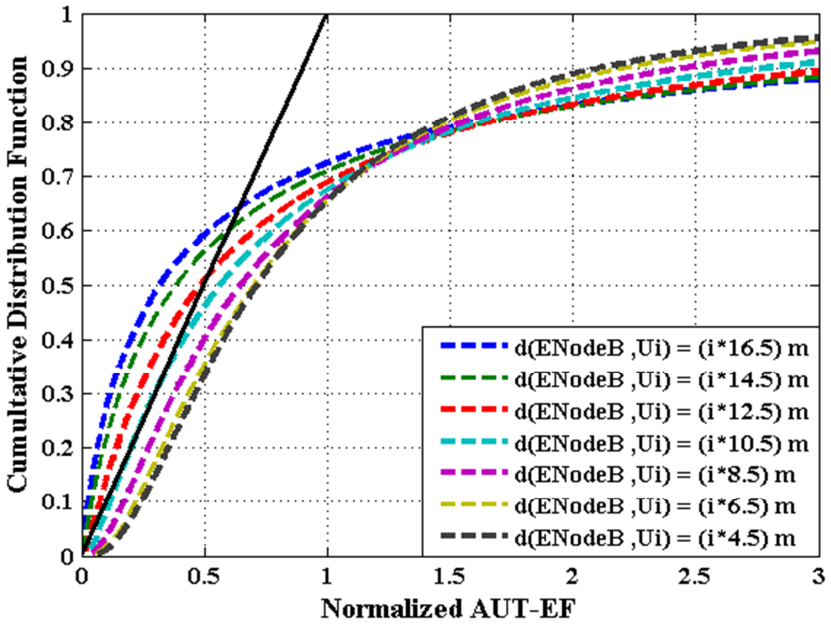


Fig. 6.3 CDF with NAUT-EMF and $(\alpha - \text{var}, \beta = 1)$ GPF Parameterization $(\beta_{\bar{T}} = 0.01)$ for 60 users scenario being equally distributed from the eNodeB base station to the edge of cell under uniform power allocation and FDD downlink transmission with a system bandwidth of 20MHz

oblique black line represents the NGMN fairness requirement in the CDF domain. Based on Eq. 6.3, any normalized throughput distribution has to lie on the right side of the NGMN requirement without crossing the black line. When the distance is $d\left(eNB, UE_i\right) \leq 10.5$ and $\alpha = 0.3$, the considered scenarios respect the NGMN fairness requirement and the schedulers are considered to be fair from the NGMN requirement point of view. Other scenarios are declared unfair. The undesired scenarios for the set of distances $d\left(eNB, UE_i\right) = \{12.5, 14.5, 16.5\}$ are able to respect the NGMN requirement only and only if $\alpha \nearrow$ TTI-by-TTI. When $d\left(eNB, UE_i\right) = 10.5$, the CDF function distribution $\psi_i\left(\hat{\bar{T}}_i[t]\right)$ is situated at the limit when the system can be declared fair. Actually, this is the *feasible situation* since it is preferable to schedule users at each TTI in such a way that the NGMN requirement is respected and the system throughput can be improved by situating the CDF curve very close to the limit of the NGMN requirement. Based on

Figures 6.2.a, 6.2.b and 6.3, the scenario when the distance is $d\left(eNB,UE_i\right)=16.5$, can be feasible and implicitly fair only and only if $\alpha > 0.5$. To conclude, *a system can move from the unfair to the fair NGMN region when the GPF-SP rule is used only and only if* $\alpha \nearrow$ *from its initial value*. The aforementioned conclusion keeps valid when the GPF-DP parameterization is performed, and the next sub-section provides the necessary explanations. The aim of the current section is to propose a set of GPF scheduling policies in order to increase the number of TTIs when the system can be declared feasible from the NGMN requirement point of view.

6.2.3 System Model for DSR-SMOO MDP Focusing on the NGMN Fairness Requirement

The main purpose of the DSR-SMOO problem focusing the NGMN fairness requirement is to find at each TTI the optimal set of fairness parameters $\left(\alpha_t^{opt},\beta_t^{opt}\right)$ in order to maintain a feasible region in the CDF domain as long as possible. The optimization problem being focused on the fairness performance from Eq. 3.35 is reloaded here for the particular case of GPF-DP such that:

$$\begin{aligned}\left(P_F\right): & \max_{\pi_{RB}[t]} \sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|} b_{i,j}[t]\cdot\frac{\left(r_{i,j}[t]\right)^{\beta_t}}{\left(\overline{T}_i[t]\right)^{\alpha_t}} \\ \left(C_F\right)\ & s.t.\ \sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t]=1,\ \forall j\in\mathcal{B} \\ & b_{i,j}[t]\in\{0,1\},\ \forall i\in\mathcal{U}_t,\forall j\in\mathcal{B}\end{aligned} \tag{6.5.a}$$

where $\left(P_F\right)$ is the optimization problem for a given set of fairness parameters $\left(\alpha_t,\beta_t\right)$ at TTI t and $\left(C_F\right)$ is the set of constraints which indicates that at each TTI t only one user $\forall i\in\mathcal{U}_t$ can be assigned to a single resource block $\forall j\in\mathcal{B}$. The performance of fairness parameters $\left(\alpha_t,\beta_t\right)$ is evaluated in terms of the NGMN criterion based on the following objective function condition:

$$\left(O_F\right):\ \psi\left(\widehat{T}_i[t]\right)\le\widehat{T}_i[t],\forall i\in\mathcal{U}_t \tag{6.5.b}$$

As mentioned in Chapter 3, based on the GPF parameterization, the DSR-SMOO problem is divided in two categories:

1. **Simple Parameterization** $\left(\alpha_t \in [0,\infty), \beta_t = 1\right)$: the scheduler controller optimizes only α_t at each TTI *t* by using CACLA1 learning with one continuous action or other discrete RL algorithms defined in Chapter 5.
2. **Double Parameterization** $\left((\alpha_t, \beta_t) \in [0,1]\right)$ where both parameters have to be optimized by using CACLA2 learning with two continuous actions or other RL approaches when a proper set of fairness parameters is mapped from the discrete controller actions.

The SMOO decision variable $c_{o,w_o}[t]$ is obtained through MUTI processing from the RL output actions $\mathcal{A}_t^{a,F}$ being focused on the NGMN fairness objective. The generalized decision vector is $c_{2,w_2}[t] = \{2, \Delta\alpha_t, \Delta\beta_t\}$, where the first element indicates the fairness objective from the overall pool of scheduling objectives and $\{\Delta\alpha_t, \Delta\beta_t\}$ is the set of parameter steps at TTI *t*, which enhances the scheduler in exploring the aggregate state space. Consequently, the GPF parameters are determined at each TTI based on Equations 6.6.a and 6.6.b. For the GPF-SP case, only Eq. 6.6.a is considered, and consequently $\left(\beta_t = 1, \Delta\beta_t = 0\right)$.

$$\alpha_t = \alpha_{t-1} + \Delta\alpha_t \tag{6.6.a}$$

$$\beta_t = \beta_{t-1} + \Delta\beta_t \tag{6.6.b}$$

The LTE controller actions $\mathcal{A}_t^{a,F}$ are taken based on the aggregate controller state space. Let us define $\mathcal{S}_t^{C,FSP}$ and $\mathcal{S}_t^{C,FDP}$ the controller state space for DSR-SMOO with simple and double parameterization, respectively. Thus the proposed state spaces take the forms of Equations 6.7.a and 6.7.b:

$$\mathcal{S}_t^{C,FSP} = \left\{\alpha_{t-1}, N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t, \mu_{\mathcal{T}}^t, \sigma_{\mathcal{T}}^t, N_U^t\right\} \tag{6.7.a}$$

$$\mathcal{S}_t^{C,FDP} = \left\{\alpha_{t-1}, \beta_{t-1}, N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t, \mu_{\mathcal{T}}^t, \sigma_{\mathcal{T}}^t, N_U^t\right\} \tag{6.7.b}$$

where $\left\{N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t\right\}$ are the elements which compute the regressed CQI state

space and the specifications are presented in Section 4.6 from Chapter 4. Based on the AUT-MMF or AUT-EMF observations, the controller state space considers the mean μ_T^t and standard deviation σ_T^t values for the log-normal distribution of user throughputs calculated by using Equations 4.6 and 4.7. The number of active users averaged over the maximum number of users (normalized value) N_U^t is considered in order to enhance the convergence of the optimal policy.

The main idea of the scheduler controller is to train the MLPNN functions for each discrete/continuous action for each given controller state $\mathcal{S}_t^{C,FSP}$ (or $\mathcal{S}_t^{C,FDP}$) in order to minimize the output errors. The set of controllable parameters $\left\{\alpha_t, \beta_t, \mu_T^t, \sigma_T^t\right\}$, which evolves based on $\left\{\Delta\alpha_t, \Delta\beta_t\right\}$ action set, indicates the controllable information. When the NGMN fairness optimality (feasibility) is reached, due to the set of uncontrollable parameters $\left\{N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t, N_U^t\right\}$, the optimal policy should act in such a way to maintain the scheduler feasibility as long as possible. The optimality region is reached based on the exploration procedure and based on the scheduler reward functions.

The proposed NGMN fairness requirement based on the reward function is analyzed by considering the test case scenario from Fig. 6.1 for the particular case of $d\left(eNB, UE_i\right) = 16.5$. By performing the Q-learning algorithm for the GPF-SP scheduling rule, the optimal set of NGMN parameters $\left(\alpha_t^{opt}\right)$ is obtained. Without going through more precise details at this stage, Fig. 6.4 shows the behavior of other scheduling approaches such as MaxTh $\left(\alpha_t = 0, \beta_t = 1\right)$, PF $\left(\alpha_t = 1, \beta_t = 1\right)$ and MaxFair $\left(\alpha_t = 1, \beta_t = 0\right)$. The MaxTh CDF curve crosses the continuous oblique line which involves the unfair character of this approach. The optimal policy and other scheduling rules assure the system convergence to the fairness region. As mentioned earlier, the idea is to define a feasible region in the CDF domain in which the CDF curve should lie on the right side and as close as possible to the fairness requirement at each TTI. The *feasible zone* defines the area between the NGMN requirement and the superior limit in the over-fairness area in which the system can be declared feasible (doted black oblique line). In

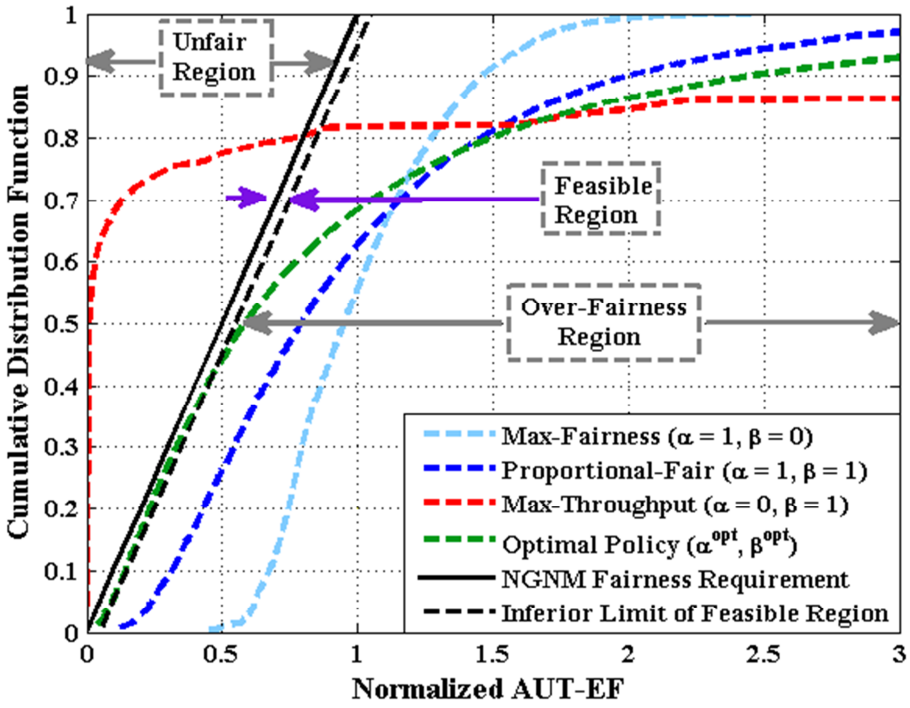


Fig.6.4 CDF with NAUT-EMF (Qualitative Tradeoff Representation) and $(\alpha-\text{var},\beta=1)$ GPF Parameterization $(\beta_{\bar{T}}=0.01)$ for 60 users scenario being equally distributed from the eNodeB base station to the edge of the cell $d(eNB,UE_i)=16.5$ under uniform power allocation and FDD downlink transmission with a system bandwidth of 20MHz.

this sense, the aggregate fairness objective function which evaluates the CDF performance against the NGMN requirement becomes:

$$\Phi\left(\widehat{\mathcal{T}}[t]\right)=\left(1/|\mathcal{U}_t|\right)\cdot\sum_{i=1}^{|\mathcal{U}_t|}\left[\left|\psi_i^{Req}\left(\widehat{\mathcal{T}}_i[t]\right)-\psi_i\left(\widehat{\mathcal{T}}_i[t]\right)\right|\right] \tag{6.8}$$

where $\Phi\left(\widehat{\mathcal{T}}[t]\right)$ has to be minimized at each TTI in order to reach the feasibility area. The NGMN fairness requirement $\psi_i^{Req}\left(\widehat{\mathcal{T}}_i[t]\right)$ can be expressed as follows:

$$\psi_i^{Req}\left(\widehat{\mathcal{T}}_i[t]\right)=\begin{cases}\widehat{\mathcal{T}}_i[t], & if\ \ \widehat{\mathcal{T}}_i[t]<1\\ 1, & if\ \ \widehat{\mathcal{T}}_i[t]\geq 1\end{cases} \tag{6.9}$$

Similarly, the feasible superior limit of the calculated percentile for observation $\widehat{\mathcal{T}}_i[t]$ and for each active user $i\in\mathcal{U}_t$ can be expressed as:

$$\psi_i^{FS}\left(\widehat{T}_i[t]\right)=\begin{cases}\widehat{T}_i[t]-\xi,\ if\ \widehat{T}_i[t]<1+\xi\\ 1,\ if\ \widehat{T}_i[t]\geq 1+\xi\end{cases} \tag{6.10}$$

where $\xi\in\mathbb{R}^{+}_{[0,1]}$ is the feasible confidence interval. When parameter ξ is small enough, the reward function is very restrictive and the obtained MDP problem may not be episodic. When ξ is large, the reward function is more flexible and the optimality region can be reached for many times by improving in this way the learning procedure. In the latter case, the user fairness performance may be improved, degrading at the same time the overall cell throughput.

Based on Eq. 6.9 and Eq. 6.10, the normalized observation $\widehat{T}_i[t]$ is feasible if and only if the following condition is respected:

$$\psi_i^{Req}\left(\widehat{T}_i[t]\right)\geq\psi_i\left(\widehat{T}_i[t]\right)\geq\psi_i^{FS}\left(\widehat{T}_i[t]\right) \tag{6.11}$$

Let us define the distance in the CDF domain $d_{i,r}^{CDF,t}=\psi_i^{Req}\left(\widehat{T}_i[t]\right)-\psi_i\left(\widehat{T}_i[t]\right)$ between the calculated percentile $\psi_i\left(\widehat{T}_i[t]\right)$ for observation $\widehat{T}_i[t]$ and its NGMN requirement $\psi_i^{Req}\left(\widehat{T}_i[t]\right)$. Then, Equation 6.11 can be rewritten as follows:

$$0\leq d_{i,r}^{CDF,t}\leq\psi_i^{Req}\left(\widehat{T}_i[t]\right)-\psi_i^{FS}\left(\widehat{T}_i[t]\right) \tag{6.12}$$

In the current approach, it is considered that if the percentile is $\psi_i\left(\widehat{T}_i[t]\right)<0.2$ based on the current distribution, then user $i\in\mathcal{U}_t$ is considered to be in *outage*. On the other side, when the percentile of user $i\in\mathcal{U}_t$ is $\psi_i\left(\widehat{T}_i[t]\right)>0.7$, the observations are situated outside of the interest domain that can influence the decision of the feasibility region. Therefore, the CDF domain in which each observation can be feasible TTI-by-TTI is $\psi_i\left(\widehat{T}_i[t]\right)\in[0.2,0.7],\ \forall i\in\mathcal{U}_t$. The same performance range is considered in [82] and for reasons of the comparison eligibility, the set of simulation results obtained in this chapter and Chapter 7 considers the same interval in the CDF domain.

Let us define the fair area in the CDF domain for all possible controller states such as $\mathcal{S}^{C,F} \in \mathcal{F}$, where $\mathcal{F} \in \mathbb{R}_{[-1,1]}^{|\mathcal{S}^C|}$. Basically, the fair zone can be divided in two sub-regions such as: $\{\mathcal{F}\} \in \{\mathcal{FAF}\} \bigcup \{\mathcal{OFF}\}$ where $\mathcal{FAF} \in \mathbb{R}_{[-1,1]}^{|\mathcal{S}^C|}$ represents the feasible or optimal region for the NGMN fairness objective. Then, the necessary and sufficient condition for $\mathcal{S}_t^{C,F} \in \mathcal{F}$ is denoted by Eq. 6.13:

$$\mathcal{S}_t^{C,F} \in \{\mathcal{F}\} \Leftrightarrow \begin{cases} 0.2 \le \psi_i\left(\widehat{\mathcal{T}}_i[t]\right) \le 0.7, \ i = 1,..,|\mathcal{U}_t| \\ d_{i,r}^{CDF,t} \ge 0, \ i = 1,..,|\mathcal{U}_t| \end{cases} \tag{6.13}$$

If $\exists d_{i,r}^{CDF,t} < 0 \ \ \forall i = 1,..,|\mathcal{U}_t|$, the system is considered unfair and $\mathcal{S}_t^{C,F} \in \mathcal{UFF}$, where $\mathcal{UFF} \in \mathbb{R}_{[-1,1]}^{|\mathcal{S}^C|}$ represents the collection of multi-dimensional data points for the unfair region. If $\exists d_{i,r}^{CDF,t} > \psi_i^{Req}\left(\widehat{\mathcal{T}}_i[t]\right) - \psi_i^{FS}\left(\widehat{\mathcal{T}}_i[t]\right), \ \forall i = 1,..,|\mathcal{U}_t|$, then the system is considered to be over-fair and $\mathcal{S}_t^{C,F} \in \mathcal{OFF}$, where $\mathcal{OFF} \in \mathbb{R}_{[-1,1]}^{|\mathcal{S}^C|}$. When the controller state space is located in the fair area, the scheduler is more interested in finding the minimum distance between each calculated percentile and its corresponding NGMN requirement such that $d_{min,r}^{CDF,t} = \underset{i \in \mathcal{U}_t}{min}\left(d_{i,r}^{CDF,t}\right)$. Based on $d_{min,r}^{CDF,t}$, the controller state is decided to belong to one of the divided sub-regions as expressed by Eq. 6.14 when $\psi_i \in [0.2; 0,7]$:

$$\mathcal{S}_t^C \in \begin{cases} \{\mathcal{UFF}\}, \ if \ \exists d_{i,r}^{CDF,t} < 0, \forall i = 1,..,|\mathcal{U}_t| \\ \{\mathcal{FAF}\}, \ if \ 0 \le d_{min,r}^{CDF,t} \le \psi_i^{Req} - \psi_i^{FS}, i = 1,..,|\mathcal{U}_t| \\ \{\mathcal{OFF}\}, \ if \ d_{min,r}^{CDF,t} > \psi_i^{Req} - \psi_i^{FS}, \forall i = 1,..,|\mathcal{U}_t| \end{cases} \tag{6.14}$$

The purpose of RL algorithms is to find the feasible state $\left(\mathcal{S}_t^C \in \mathcal{FAF}\right)$ based on the controller action $\mathcal{A}_t^{a,F}$ being applied at TTI *t* and to keep this desirable state as long as possible. Other regions such as $\{\mathcal{UFF}\}$ or $\{\mathcal{OFF}\}$ are considered undesirable for the learning procedure $\left(\mathcal{S}_t^{C,F} \notin \{\mathcal{UFF}, \mathcal{OFF}\}\right)$ when the DSR-SMOO MDP problems focusing on the NGMN fairness objective are considered.

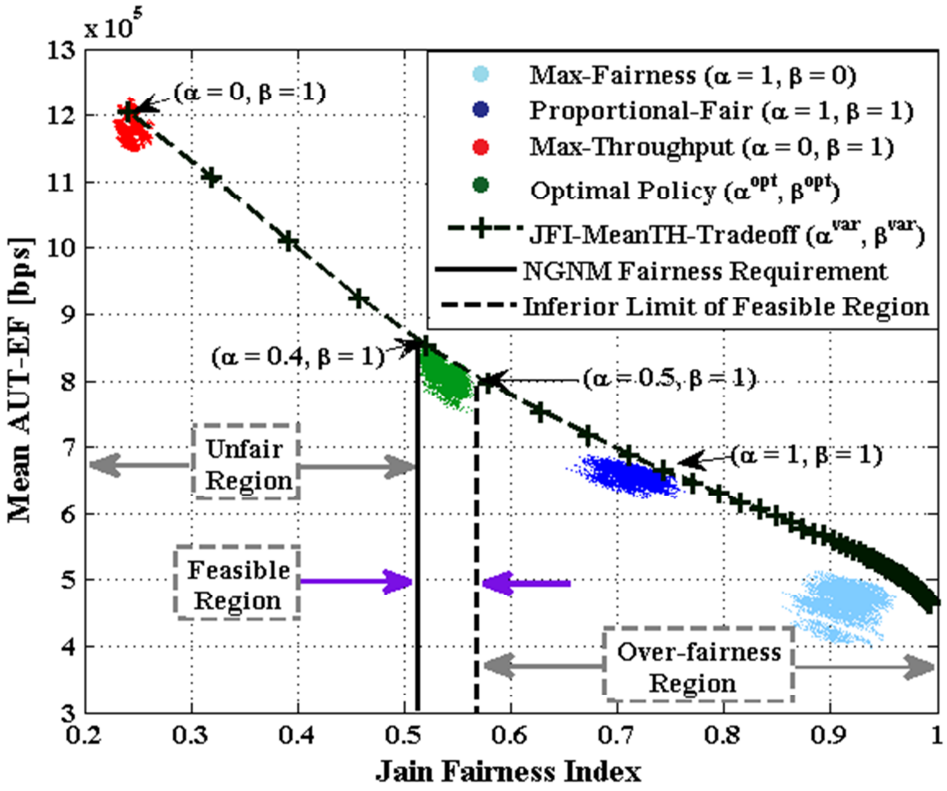


Fig.6.5 JFI-EF and Mean AUT-EF (Quantitative Tradeoff Representation) and $(\alpha - \text{var}, \beta = 1)$ GPF Parameterization $(\beta_T = 0.01)$ for 60 users scenario being equally distributed from ENodeB base station to the edge of cell $d(eNB, UE_i) = 16.5$ under uniform power allocation and FDD downlink transmission with the bandwidth of 20MHz.

In conclusion, *the qualitative representation of the fairness throughput tradeoff is used to decide the controller state feasibility TTI-by-TTI.*

The controller state status $(\mathcal{S}_t^{C,F} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\})$ decision is very important for the NGMN reward function computation. In general, the reward function should be obtained based on the objective function (e.g. Eq. 6.8). As explained in Chapter 5, the controller may choose a different time scale (multiple TTIs) in order to stabilize the objective function when the controller actions use large $(\Delta\alpha_t, \Delta\beta_t)$ steps. In order to avoid the usage of Eq. 6.8 as a reward function, *the GPF parameterization techniques should be analyzed in close collaboration with the controller state space status TTI-by-TTI.* In this sense, the JFI-EMF/AUT-EMF trade-off (Fig. 6.5) transposes Figure 6.4 in a more perceptible

representation from the parameters set (α_t, β_t) point of view. It is the case of quantitative tradeoff evaluation since a precise numerical value is depicted for all possible combinations of (α_t, β_t). The feasible zone is located in the range of $\alpha_t \in (0.4, 0.5)$ when $\beta_t = 1$. The dot star line represents the maximum mean-AUT-EMF which can be achieved for any combination of (α_t, β_t) parameters when the scenario of $d(eNB, UE_i) = 16.5$ is considered. Thus, the role of the LTE scheduler controller is to explore the trade-off curve for infinite number of state-action pairs in order to localize the fairness feasibility state.

One way to praise the controller actions is to set the reward function based on the calculated percentile to the NGMN requirement. When $\mathcal{S}_t^{C,F} \in \mathcal{F}$, the minimum distance $d_{min,r}^{CDF,t}$ is considered. When $\mathcal{S}_t^{C,F} \in \mathcal{UFF}$, the maximum distance should be calculated such that $d_{max,r}^{CDF,t} = \underset{ii \in \mathcal{U}_t}{max}\left(d_{ii,r}^{CDF,t}\right),\ \forall ii = 1,..,|\mathcal{U}_t|$ where ii represents the bearer index for which the CDF difference is $d_{ii,r}^{CDF,t} < 0$. Therefore, the reward function can be calculated by using Eq. 6.15:

$$\mathcal{RW}_t^F\left(\mathcal{S}_{t-1}^{C,F}, \mathcal{A}_{t-1}^{a,F}\right) = \begin{cases} -d_{max,r}^{CDF,t},\ \mathcal{S}_{t-1}^{C,F} \in \mathcal{UFF} \\ 1 - d_{min,r}^{CDF,t},\ \mathcal{S}_{t-1}^{C,F} \in \mathcal{F} \end{cases} \tag{6.15}$$

The reward representation from Eq. 6.15 has a big disadvantage. If the controller explores the feasible region from Fig. 6.5 and based on the mapped scheduling decision, the fairness parameter step becomes $\Delta\alpha_t = -0.5$, the scheduler moves to $\alpha_{t+1} \in (0, 0.1)$ in the next state which is known to represent the unfair region. But the reward $\mathcal{RW}_{t+1}^F\left(\mathcal{S}_t^{C,F}, \mathcal{A}_t^{a,F}\right)$ remains very high due to the averaging effect of observation $\widehat{T}_i[t]$. In this case, *the scheduler reward is noisy*. Then, the scheduler and the controller should perform at different time scales as shown in Sub-section 5.6.7 in order to stabilize the noisy effect. In order to avoid such complicated architecture, *the presence of* (α_t, β_t) *from the controller state space should be exploited for the reward function computation*. In conclusion, *the quantitative representation is preferred in the detriment of qualitative tradeoff for*

the fairness reward function computation. The qualitative evaluation can be used in order to localize the feasible, over-fair or unfair states and the reward function is computed based on the fairness parameters.

In order to enhance the convergence to the feasible region, the minimum/maximum distances can be inserted in the controller state space for both parameterization schemes as follows:

$$\mathcal{S}_t^{C,FSP} = \begin{cases} \left[\alpha_{t-1}, N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t, \mu_{\mathcal{T}}^t, \sigma_{\mathcal{T}}^t, -d_{max,r}^{CDF,t}, N_U^t\right], \ if \ \mathcal{S}_t^{C,FSP} \in \mathcal{UFF} \\ \left[\alpha_{t-1}, N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t, \mu_{\mathcal{T}}^t, \sigma_{\mathcal{T}}^t, d_{min,r}^{CDF,t} N_U^t\right], \ if \ \mathcal{S}_t^{C,FSP} \in \{\mathcal{FAF}, \mathcal{OFF}\} \end{cases} \tag{6.16.a}$$

$$\mathcal{S}_t^{C,FDP} = \begin{cases} \left[\alpha_{t-1}, \beta_{t-1}, N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t, \mu_{\mathcal{T}}^t, \sigma_{\mathcal{T}}^t, -d_{max,r}^{CDF,t}, N_U^t\right], \ if \ \mathcal{S}_t^{C,FSP} \in \mathcal{UFF} \\ \left[\alpha_{t-1}, \beta_{t-1}, N_{CL}^A, \sigma_{CL}^t, ks_t, d_{ks}^t, \mu_{\mathcal{T}}^t, \sigma_{\mathcal{T}}^t, d_{min,r}^{CDF,t} N_U^t\right], \ if \ \mathcal{S}_t^{C,FSP} \in \{\mathcal{FAF}, \mathcal{OFF}\} \end{cases} \tag{6.16.b}$$

Based on Eq. 6.16.a and Eq. 6.16.b, the obtained state space consists a precise information of how far or close the system is from the feasible state when $(\alpha_{t-1}, \beta_{t-1})$ has been applied in the previous state.

When the scheduler is over-fair (Fig. 6.5), any increase of α_t moves the scheduler further away from the optimal region in the over-fairness CDF region. On the other pole, when the scheduler is unfair, it is undesirable to decrease α_t parameter due to the fact that the scheduler moves further in the unfair CDF region. Therefore, for the GPF-SP case, when $\mathcal{S}_t^{C,F} \in \mathcal{UFF}$, then $\alpha_t \nearrow$, and when $\mathcal{S}_t^{C,F} \in \mathcal{OFF}$, then $\alpha_t \searrow$. Based on the aforementioned characteristics, the reward function for the simple parameterization case becomes:

$$\mathcal{RW}_t^{FSP} = \begin{cases} \alpha_{t-1} - \alpha_{t-2}, \ if \, \alpha_{t-1} \geq \alpha_{t-2}, \mathcal{S}_{t-1}^{C,FSP} \in \mathcal{UFF}, \mathcal{S}_t^{C,FSP} \in \mathcal{UFF} \\ -1, \ if \, \alpha_{t-1} < \alpha_{t-2}, \mathcal{S}_{t-1}^{C,FSP} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}, \mathcal{S}_t^{C,FSP} \in \mathcal{UFF} \\ 0, \ if \, \alpha_{t-1} > \alpha_{t-2}, \mathcal{S}_{t-1}^{C,FSP} \in \mathcal{UFF}, \mathcal{S}_t^{C,FSP} \in \mathcal{OFF} \\ 1, \ if \ \mathcal{S}_{t-1}^{C,FSP} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}, \mathcal{S}_t^{C,FSP} \in \mathcal{FAF} \\ -1, \ if \, \alpha_{t-1} \geq \alpha_{t-2}, \mathcal{S}_{t-1}^{C,FSP} \in \{\mathcal{FAF}, \mathcal{OFF}\}, \mathcal{S}_t^{C,FSP} \in \mathcal{OFF} \\ \alpha_{t-1} - \alpha_{t-2}, \ if \, \alpha_{t-1} < \alpha_{t-2}, \mathcal{S}_{t-1}^{C,FSP} \in \mathcal{OFF}, \mathcal{S}_t^{C,FSP} \in \mathcal{OFF} \end{cases} \tag{6.17}$$

As indicated by Eq. 6.17, the proposed reward function considers 11 types of transitions in which the feasible state is granted with the highest reward value. That is, the MLPNN function is trained based on the stochastic gradient descent principle which aims to minimize the error between the target state-action value $Q_t^T\left(\mathcal{S}_{t-1}^{C,FSP},\mathcal{A}_{t-1}^{a,FSP}\right)=\mathcal{RW}_t^{FSP}\left(\mathcal{S}_{t-1}^{C,FSP},\mathcal{A}_{t-1}^{a,FSP}\right)+\gamma\cdot Q_t^F\left(\mathcal{S}_{t-1}^{C,FSP},\mathcal{A}_{t-1}^{a,FSP}\right)$ and the known forwarded value $Q_t^F\left(\mathcal{S}_{t-1}^{C,FSP},\mathcal{A}_{t-1}^{a,FSP}\right)$. The optimal controller policy aims to select the best mapped action $\Delta\alpha_t^{opt}$ to be performed in the current TTI. The main disadvantage of DSR-SMOO MDP with the GPF-SP parameterization is the fact that $\alpha_t\in[0,\infty)$ should be adapted in a very large domain and this may take too much time to revenue to the optimal value when the scheduler conditions change drastically. For this reason, the fairness optimization based on GPF-DP can offer better performances since it has to adapt two-parameter steps which are able to find the optimal state much faster when the radio conditions seriously fluctuate.

From the viewpoint of the controller functionality, the proposed approach refers to CACLA-learning which makes use of two continuous decisions (CACLA2) in terms of $\left(\Delta\alpha_t,\Delta\beta_t\right)$. This means that both fairness parameters should be included in the reward function since the optimal state depends on the evolution of both parameters as indicated by Eq. 6.16.b. Another advantage of CACLA2 when compared against CACLA1 or other RL approaches with single action is the presence of multiple $\left(\alpha_t^{opt},\beta_t^{opt}\right)$ optimal solutions for the NGMN fairness requirement. When $\mathcal{S}_t^{C,FSP}\in\mathcal{FAF}$, the parameter set $\left(\alpha_t^{opt},1\right)$ is unique, whereas $\mathcal{S}_t^{C,FDP}\in\mathcal{FAF}$ reveals the existence of multiple optimal solutions $\left(\alpha_t^{opt},\beta_t^{opt}\right)$. In Figure 6.5, the feasibility $\mathcal{S}_t^{C,FSP}\in\mathcal{FAF}$ is reached when, let us say, $\left(\alpha_t^{opt}=0.5,\beta_t^{opt}=1\right)$. If the fairness parameters consider $\left(\alpha_t=\alpha^{opt},\beta_t\searrow\right)$, then $\mathcal{S}_t^{C,FDP}\in\mathcal{UFF}$ and if $\left(\alpha_t\nearrow,\beta_t=1\right)$, then the system tends to become over-fair $\mathcal{S}_t^{C,FDP}\in\mathcal{OFF}$. Based on these principles, the state feasibility can be reached for multiple optimal sets of fairness parameters when $\left(\alpha_t\searrow,\beta_t\searrow\right)$. Then, the reward function for SMOO-GPF-DP can be divided as indicated in Eq. 6.18:

$$\mathcal{RW}_t^{FDP} = \begin{cases} \mathcal{RW}_t^{UFDP}, \ if \ \mathcal{S}_{t-1}^{C,FDP} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}, \mathcal{S}_t^{C,FDP} \in \mathcal{UFF} \\ 1, \ if \ \mathcal{S}_{t-1}^{C,FDP} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}, \mathcal{S}_t^{C,FDP} \in \mathcal{FAF} \\ \mathcal{RW}_t^{OFDP}, \ if \ \mathcal{S}_{t-1}^{C,FDP} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}, \mathcal{S}_t^{C,FDP} \in \mathcal{OFF} \end{cases} \quad (6.18)$$

The reward function $\mathcal{RW}_t^{UFDP}$, when the current state is unfair $\mathcal{S}_t^{C,FDP} \in \mathcal{UFF}$, can be modeled by using the tuple $(\Delta\alpha_{t-1}, \Delta\beta_{t-1}, \alpha_{t-1}, \beta_{t-1})$, where the fairness steps are $\Delta\alpha_{t-1} = \alpha_{t-1} - \alpha_{t-2}$ and $\Delta\beta_{t-1} = \beta_{t-1} - \beta_{t-2}$. In this case, the parameters should respect the conditions of $\Delta\alpha_{t-1} > 0$ and $\Delta\beta_{t-1} < 0$ in order to move the system from the unfair region. Then, the decision can be further divided in two situations: when $\alpha_{t-1} < \beta_{t-1}$, the reward function should contain a weighted sum of $(\Delta\alpha_{t-1}, \Delta\beta_{t-1})$, and when $\alpha_{t-1} > \beta_{t-1}$, the role of β_{t-1} is not important anymore and the reward should focus only on $\Delta\alpha_{t-1}$. In the opposite case, when $\Delta\alpha_{t-1} < 0$ and $\Delta\beta_{t-1} > 0$, the action should be severely punished. Based on principles exposed above, the reward function for the unfair zone is calculated based on Eq. 6.19.a.

$$\mathcal{RW}_t^{UFDP} = \begin{cases} 0.5 \cdot (|\Delta\alpha_{t-1}| + |\Delta\beta_{t-1}|), if \ \Delta\alpha_{t-1} > 0, \Delta\beta_{t-1} < 0, \alpha_{t-1} < \beta_{t-1} \\ -0.5 \cdot (1 + |\Delta\beta_{t-1}|), if \ \Delta\alpha_{t-1} < 0, \Delta\beta_{t-1} < 0, \alpha_{t-1} < \beta_{t-1} \\ -0.5 \cdot (|\Delta\alpha_{t-1}| + 1), if \ \Delta\alpha_{t-1} > 0, \Delta\beta_{t-1} > 0, \alpha_{t-1} < \beta_{t-1} \\ |\Delta\alpha_{t-1}|, if \ \Delta\alpha_{t-1} > 0, \Delta\beta_{t-1} < 0, \alpha_{t-1} > \beta_{t-1} \\ -1, if \ \Delta\alpha_{t-1} < 0, \Delta\beta_{t-1} > 0 \end{cases} \quad (6.19.a)$$

For the over-fairness case, the reward function should follow the opposite direction of Eq. 6.19.a. Equation 6.19.b. highlights the proposed reward model for the particular situation when the current controller state is $\mathcal{S}_t^{C,FDP} \in \mathcal{OFF}$. The desirable situation is denoted by $\{\Delta\alpha_{t-1} < 0, \Delta\beta_{t-1} > 0\}$ and the undesirable one by

$$\mathcal{RW}_t^{OFDP} = \begin{cases} 0.5 \cdot (|\Delta\alpha_{t-1}| + |\Delta\beta_{t-1}|), if \ \Delta\alpha_{t-1} < 0, \Delta\beta_{t-1} > 0, \alpha_{t-1} > \beta_{t-1} \\ -0.5 \cdot (1 + |\Delta\beta_{t-1}|), if \ \Delta\alpha_{t-1} > 0, \Delta\beta_{t-1} < 0, \alpha_{t-1} > \beta_{t-1} \\ -0.5 \cdot (|\Delta\alpha_{t-1}| + 1), if \ \Delta\alpha_{t-1} < 0, \Delta\beta_{t-1} < 0, \alpha_{t-1} > \beta_{t-1} \\ |\Delta\alpha_{t-1}|, if \ \Delta\alpha_{t-1} < 0, \Delta\beta_{t-1} > 0, \alpha_{t-1} < \beta_{t-1} \\ -1, if \ \Delta\alpha_{t-1} > 0, \Delta\beta_{t-1} < 0 \end{cases} \quad (6.19.b)$$

the case when $\{\Delta\alpha_{t-1} > 0, \Delta\beta_{t-1} < 0\}$. In the first situation, α_{t-1} and β_{t-1} must be compared in order to determine whether or not $\Delta\beta_{t-1}$ can help in reaching the feasible scheduler state.

The role of the analyzed RL approaches is to explore different combinations of JFI-Mean-AUT tradeoff levels for infinite number of controller states in order to converge to the feasible state. The non-actor based RL schemes attract a disadvantage due to their ability to explore such curves from Fig. 6.5 without having well defined the state-action values in order to determine whether or not the applied set of continuous actions $(\Delta\alpha_t, \Delta\beta_t)$ represents good solutions to approximate a given controller state. The usage of the actor-critic schemes has the advantage that even in the exploration stage, the algorithms are capable to localize the optimal region and to spend more time on that zone without wasting the exploration time on irrelevant controller state space regions.

6.2.4 Comparative Methods

Based on the proposed controller state space, the parameterization of the GPF scheduling schemes is achieved by using the instantaneous aggregate CQI information, the minimum/maximum distance from the calculated percentiles and the NGMN fairness requirement. By using the considered state space, the MLPNN fairness functions are learned through exploration and/or experience replay stages in order to be exploited for the real time scheduling processes.

The performance of the learned functions are compared and matched against the existing scheduling approaches focusing on adaptive fairness-throughput tradeoff [79], [82]. Both of the existing techniques use the GPF-SP scheme where $\Delta\alpha_t$ parameter step can be decided based on the quantitative tradeoff evaluation [79] or on the qualitative representation (CDF domain) [82].

In the first approach, the expectation of the predicted instantaneous throughput is calculated at each TTI, and $\Delta\alpha_t$ is calculated based on the

difference between the expected JFI and the JFI constraint [79]. This technique is entitled *Maximizing Throughput with adjustable fairness* (MT). In order to enhance the comparative analysis, the same scheme is used but the step parameter is determined based on the minimum or maximum distances in the CDF domain as expressed by the following equation:

$$\Delta\alpha_t = \begin{cases} \max\limits_{i\in\mathcal{U}_t}\left\{ d_i^t\left[\psi_i^{Req}\left(\underrightarrow{\widehat{T}_i}[t]\right), \psi_i\left(\underrightarrow{\widehat{T}_i}[t]\right)\right]\right\}, & \mathcal{S}_t^{C,F} \in \mathcal{UFF} \\ -\min\limits_{i\in\mathcal{U}_t}\left\{ d_i^t\left[\psi_i^{Req}\left(\underrightarrow{\widehat{T}_i}[t]\right), \psi_i\left(\underrightarrow{\widehat{T}_i}[t]\right)\right]\right\}, & \mathcal{S}_t^{C,F} \in \{\mathcal{FAF}, \mathcal{OFF}\} \end{cases} \quad (6.20)$$

where $\underrightarrow{\widehat{T}_i}[t]$ represents the expectation of the predicted user throughput by using the exponential or the mean moving filters. More details about the throughput prediction computation can be found in [79].

The second method aims to store multiple IUT values in drops (multiple TTIs) and the adaptation is achieved at each time drop in order to reduce the computational complexity. The GPF-SP parameter is then adjusted by using Equation 6.20 by simply removing the expectation value with a set of instantaneous user throughputs such that $\underrightarrow{\widehat{T}_i} = \{T_i[t], T_i[t-1], ..., T_i[t-D_T]\}$, $\forall i \in \mathcal{U}_t$ and D_T is the fairness adaptation drop period. In order to make this proposal suitable for the TTI-by-TTI adaption, the set of IUTs is replaced by the average user throughputs and the parameter adaptation is performed based on Eq. 6.20 where $\underrightarrow{\widehat{T}_i}[t] = \widehat{T}_i[t]$. This method is known as *Adaptive Scheduling for fairness control* (AS) [82].

Even if the CQI probability mass function is considered in the expected user throughput computation [79], the major drawback of these proposals refers to the fact that the channel conditions are not considered for the $\Delta\alpha_t$ decision and computation. It is expected that much finer adaptation can be obtained if additional CQI statements are considered in the GPF parameterization techniques. In this sense, for the RL based GPF-SP/DP approaches, $(\Delta\alpha_t, \Delta\beta_t)$ steps are obtained based on the trained MLPNN structures which consider the aggregate CQI information and the distances in the CDF domain as suggested in Eq. 6.16.

Sets of sustainable scheduling policies for DSR-SMOO problems being focused on the NGMN fairness criterion are proposed in [245] and [246] where the ITU Vehicular A channel type and the Rayleigh fading model are used. In [245], the CACLA1 policy which parameterizes the GPF-SP scheduling rule outperforms from the viewpoint of the percentage of feasible TTIs other policies obtained with different RL approaches such as: ACLA, SARSA, Q and QV. Also, CACLA1 offers much better performances when compared against the existing techniques from [79] and [82]. It is important to notice that the CQI channel aggregation from Chapter 4 is not used in the controller state space computation in [245] and [246]. In Sub-section 6.2.5 it is shown that the proposed policies are not able to perform better than classical approaches such as AS or MT when the CQI aggregation module is not considered under the fast Jakes fading model. In [246], the CACLA2 policy performs much better than other scheduling policies provided by CACLA1, QVMAX, QVMAX2, QV2 and DoubleQ learning.

Extensive simulation results show that only when the CQI aggregation module proposed in Chapter 4 is used, the existing methods such as MT and AS can be outperformed. This way, the proposed RL policies are able to adapt much better the fairness parameters when compared with the existing approaches by increasing the number of TTIs when the scheduler is declared feasible and by reducing the percentage of TTIs when the scheduler is unfair. In the following, the performance of the proposed RL schemes is analyzed for the DSR-SMOO problems focusing on the NGMN fairness requirement with exponential and median moving filters for the computation of the average user observations.

6.2.5 Performance Evaluation of Sustainable Scheduling Policies Focusing on NGMN Fairness Criterion

The fluctuations in the CDF domain become significant for higher forgetting factor $\left(\beta_{\bar{T}} > 0.01\right)$ values when the AUT-EMF observations are used. On the other hand, when $\beta_{\bar{T}}$ is very low, the scheduling procedure becomes indistinguishable at every TTI. Based on extensive simulation results, the

optimum forgetting factor is $\beta_{\bar{T}} = 0.01$ which in fact permits to avoid the fluctuations of the scheduling results and to offer enough contribution in the AUT computation in order to localize the NGMN feasibility zone. This limitation becomes impracticable when other scheduling objectives require adaptive weighting factors in the AUT observations. In order to avoid the drawbacks of AUT-EMF observations, the median filter is preferred to be used due to its lower fluctuations in the scheduling results for dynamic averaging time windows. In this sub-section, the AUT-EMF computation is used in order to highlight the importance of using different CQI aggregation techniques in the optimization problems which concern the NGMN fairness objective. This sub-section is organized as follows: Sub-section 6.2.5.1 presents the simulation scenario and the parameter settings, Sub-section 6.2.5.2 analyzes the performance of scheduling rules by using the AUT-EMF observations and Sub-section 6.2.5.3 highlights the simulation results and shows the sustainability of the proposed scheduling policies for different filter lengths when the AUT-MMF observations are used.

6.2.5.1 Simulation Scenario

The DSR-SMOO focusing on the NGMN fairness requirement considers a general scenario where the number of users fluctuates in the range of [15; 120] for computational complexity reasons with a switching period of 1000 TTIs. The user speed is 120kmph with fast-fading Jakes model. The Jakes model experiences very fast fading and the impact of the CQI aggregation module with different re-assignment preprocessing schemes is analyzed in Sub-section 6.2.5.2. The CQI report is errorless, periodic and full-band in order to have complete information about the CQI statistics for a certain number of active users. The rest of physical layer parameters are imported from the 3GPP simulation scenarios [36]. The infinite traffic model is analyzed and the RLC layer is modeled by using the Transmission Mode (TM) without retransmissions due to the fact that the results are oriented more on the decision of fairness parameters for the first transmission.

Both types of fairness adaptation based scheduling rules are used in terms of GPF-SP and GPF-DP parameterizations. The scheduling policies obtained by

Table 6.1 LTE Scheduler Parameters for DSR-SMOO Focusing on NGMN Fairness

Parameters Name	Description/Values
System Bandwidth/Cell Radius	20 MHz/1000m
User Speed/Mobility Model	120 Kmph/Random Direction
Channel Model	Jakes Model
Path Loss / Penetration Loss	Macro Cell Model / 10 dB [36]
Interfered Cells/Shadowing STD	0/8 dB [36]
Carrier Frequency/DL Power	2GHz/43dBm [36]
Frame Structure	FDD [36]
CQI Reporting Mode	Full-band, periodic at each TTI
PUCCH Model	Errorless
Scheduler Type	GPF-SP/GPF-DP
Traffic Type	Infinite Buffer
RLC ARQ	Transmission Mode (no retransmissions)
AMC Levels	QPSK (1/3, 1/2, 2/3), 16-QAM (1/2, 2/3, 5/6) 64-QAM (2/3, 5/6) [36]
Target BLER	10% (Appendix B)
Number of Users $\left(\max \ \lvert\mathcal{U}\rvert\right)$	Variable : 15-120
RL Algorithms	Q-L, DoubleQ-L, SARSA, QV, QV2, QVMAX, QVMAX2, ACLA, CACLA1, CACLA2
Discrete MLPNN Actions	$\Delta\alpha = \left\{\pm 10^{-4}, \pm 10^{-3}, \pm 10^{-2}, \pm 10^{-1}, \pm 5 \cdot 10^{-2}, 0\right\}$
Continuous MLPNN Actions	$\left(\Delta\alpha, \Delta\beta\right) \in \mathbb{R}_{[-1;1]}$
Controller Timescale	1 TTI
Number of MLPNN layers / Activation Functions	3/input layer: linear activation, hidden layer: tangent hyperbolic activation, output layer: linear activation
Number of Hidden Nodes	60
Exploration/ Experience Replay/Exploitation Periods	3000/1000 (Q-L, DoubleQ-L, SARSA)/200
AUT-EMF Forgetting Factor $\left(\beta_{\bar{T}}\right)$	0.01
NGMN Confidence Factor $\left(\xi\right)$	0.05
CQI Aggregation Schemes	Variable Mass Modes $\left\{Top3, Top4, Top5\right\} : N_{CT} = \left\{64, 128, 256, 512\right\}$

using the RL algorithms are compared against the existing approaches such as MT and AS. The RL policies use both types of action spaces: discrete (Q, Double-Q, SARSA, QV, QV2, QVMAX, QVMAX2, ACLA) and continuous (CACLA1, CACLA2). For the discrete case, the following set of fairness steps is considered based on the quality of the results: $\Delta\alpha = \left\{\pm 10^{-4}, \pm 10^{-3}, \pm 10^{-2}, \pm 10^{-1}, \pm 5 \cdot 10^{-2}, 0\right\}$. The MLPNN structure makes use of three layers with the tangent hyperbolic

Table 6.2 LTE Scheduler Controller Parameters for DSR-SMOO Focusing on NGMN Fairness Requirement

RL Algorithm (Fairness SMOO)	Learning Rates for Action Values (η^{Q})	Learning Rates for State Values (η^{V})	Discount Factor (γ)	Exploration Type (ε,τ)
Q	0.001	-	0.99	Greedy $(\varepsilon=10^{-4})$
DQ	0.001	-	0.99	Greedy $(\varepsilon=10^{-4})$
SARSA	0.001	-	0.99	Boltzmann $(\tau=10)$
QV	0.01	0.0001	0.99	Boltzmann $(\tau=1)$
QV2	0.001	0.00001	0.95	Boltzmann $(\tau=1)$
QVMAX	0.01	0.0001	0.99	Boltzmann $(\tau=10)$
QVMAX2	0.001	0.00001	0.95	Boltzmann $(\tau=1)$
ACLA	0.01	0.01	0.99	Greedy $(\varepsilon=0.5)$
CACLA1	0.01	0.01	0.99	Greedy $(\varepsilon=0.5)$
CACLA2	0.01	0.01	0.99	Greedy $(\varepsilon=0.5)$

activation function for the hidden layer. Based on some simulation results, the optimum number of MLPNN hidden nodes is 60 for the DSR-SMOO problems focusing on NGMN fairness in order to avoid the over-fitting problems and to reduce the computational complexity. The neural network weights are trained by using an exploration period of 3000s for all RL approaches. In addition, the experience replay stage is used by the RL algorithms which consider only the action values such as Q-L, DoubleQ-L or SARSA. The trained MLPNN structure is exploited for 200s in order to highlight the sustainability of the learned policies being obtained with different RL algorithms. The rest of the scheduler and controller parameters is illustrated in Table 6.1 and the parameter settings of RL algorithms are presented in Table 6.2. Apart from the actor-critic schemes and Q-Learning or DoubleQ-Learning, other RL algorithms are using the Boltzmann distribution probability in order to decide the policy evaluation or the policy improvement during the exploration stage. The controller parameters are obtained based on extensive simulations and the optimum values are listed in Table 6.2.

The MLPNN weights are trained in the first instance without considering the CQI aggregation information from Chapter 4 in the controller state space by

simply removing the elements $\left\{N_{CL}^{A}, \sigma_{CL}^{t}, ks_t, d_{ks}^{t}\right\}$ from Equations 6.16.a and 6.16.b. In the second case, 12 configurations are used for the CQI aggregation schemes with Top3, Top4 and Top5 mass modes and the numbers of pre-processed CQI data centers of $N_{CT} = \{64, 128, 256, 512\}$. In this way, the implementation of the CQI aggregation techniques becomes mandatory. The number of $N_{CT} = 1024$ preprocessed CQI centers is omitted due to the very high computational complexity which is involved in the exploration stage.

6.2.5.2 DSR-SMOO MDP Based on Average Throughput Observations with Exponential Moving Filter

Figure 6.6 shows the performances of the proposed RL algorithms against the existing methods (AS and MT) from the perspective of percentage of TTIs under the scheduler states $\mathcal{S}_t^{C,F} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ when the CQI aggregation module is not considered. The percentage of TTIs are calculated based on $p_{TTI}^{stat}[\%] = \left(N_{TTI}^{stat} / N_{TTI}^{Exploit}\right) \cdot 100\%$, where N_{TTI}^{stat} is the number of TTIs when the scheduler stays unfair, feasible or over-fair. Due to the variability of these results, the obtained policies are tested by using 10 exploitations with different initial positions of the mobile users. Then, the percentage of TTIs p_{TTI}^{stat} for different scheduler states is represented by using the mean and the error (standard deviation) values. As expected, the static GPF parameterization $(\alpha = 1, \beta = 1)$ indicates the highest percentage of TTIs when the scheduler is declared over-fair. The same results are obtained for Q-L, QVMAX2 and DoubleQ policies which use the simple parameterization techniques $\left(\alpha \in \mathbb{R}_{[0,3]}^{+}, \beta = 1\right)$. The highest amount of TTIs when the system is unfair is obtained when the QV2 policy is exploited. The best performances in terms of the mean percentages of TTIs when the scheduler is feasible is denoted by CACLA2 and ACLA actor-critic schemes with simple and double GPF parameterizations, respectively. The QV-learning shows the highest variability in terms of the number of TTIs when the scheduler state is $\mathcal{S}_t^{C,FSP} \in \{\mathcal{FAF}, \mathcal{OFF}\}$. From Fig. 6.6 it can be concluded that the proposed

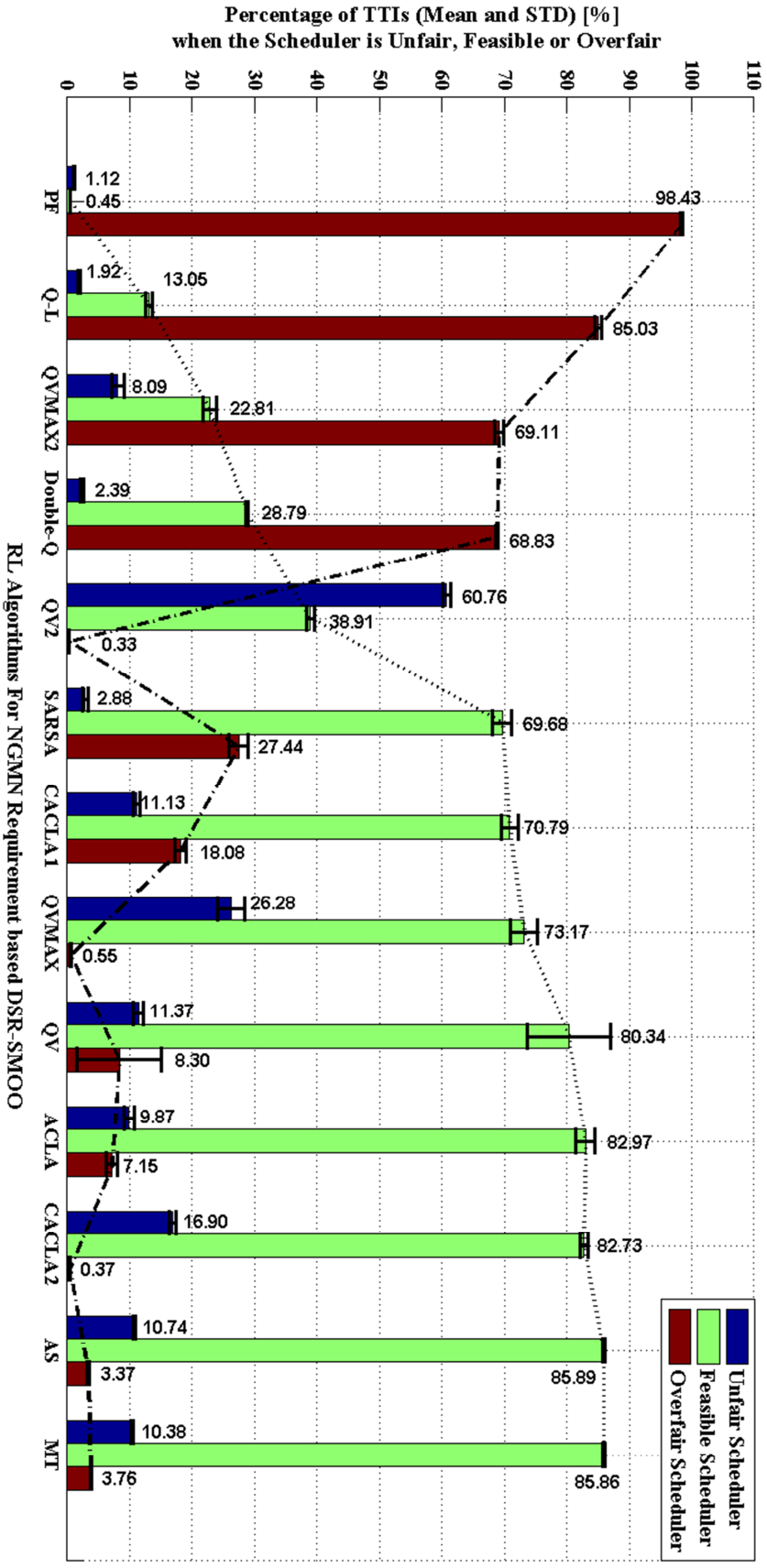


Fig. 6.6 Percentages of TTIs (Mean and Standard Deviation) Under the Unfair, Feasible or Over-fair Scheduler States (No CQI Aggregation scheme is considered)

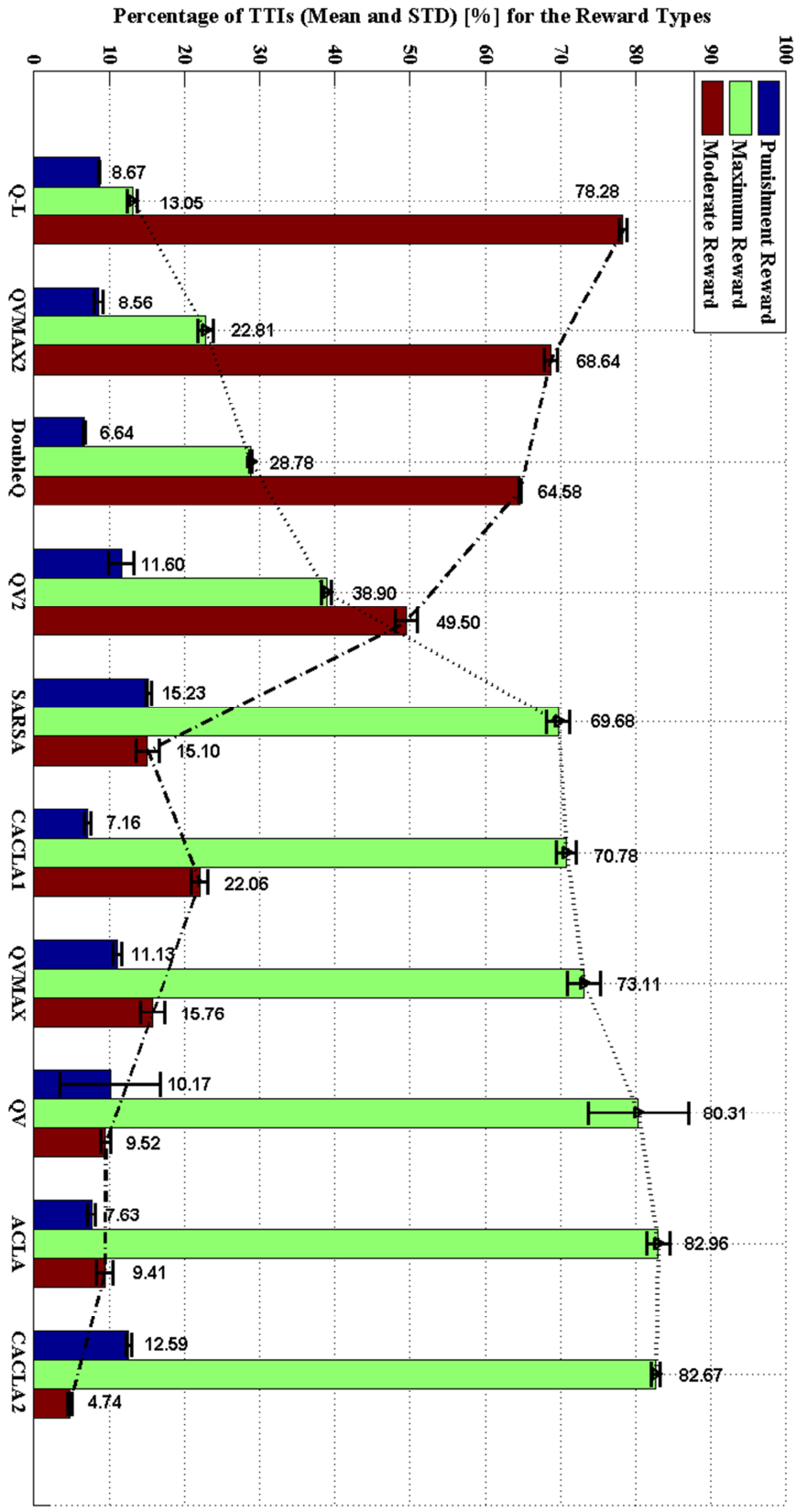


Fig. 6.7 Percentages of TTIs (Mean and Standard Deviation) for the Reward Type (No CQI Aggregation scheme is considered)

policies are not able to perform better in terms of the number of TTIs when the scheduler stays feasible than the existing methods (AS and MT) when the CQI aggregation schemes are not taken into account in the controller state space.

Figure 6.7 depicts the percentages of TTIs (mean and STD values) for different testing reward types such as punishment reward $\left(\mathcal{RW}_t^F = \{-1\}\right)$, moderate reward $\left(\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}\right)$ and the maximum reward $\left(\mathcal{RW}_t^F = \{+1\}\right)$ which in fact denotes the feasibility state. These types of rewards are used in the exploitation stage only to highlight the performance of each RL scheme. The performance is presented in a similar way to Figure 6.6, but the percentages of TTIs are replaced by the set of $\left\{\overline{p}_{TTI}^{F,PSH}, \overline{p}_{TTI}^{F,mRW}, \overline{p}_{TTI}^{F,MRW}\right\}$ where $\overline{p}_{TTI}^{F,PSH}$ is the mean percentage of TTIs when the reward is punishment, $\overline{p}_{TTI}^{F,mRW}$ represents the mean percentage of TTIs when the reward is moderate, and $\overline{p}_{TTI}^{F,MRW}$ is the mean percentage of TTIs when the maximum reward is received. The highest amount of moderate rewards is obtained when Q-L, QVMAX2, DoubleQ and QV2 learning procedures are used denoting in fact the difficulty for these policies in reaching the feasible states. The experience replay stage is not able to improve the quality of the learned policy for the Q-L and DoubleQ learning procedures. A slight improvement can be detected in the SARSA scheduling policy which can achieve about $\overline{p}_{TTI}^{F,MRW} = 70\%$ maximum rewards and $\overline{p}_{TTI}^{F,mRW} = 15\%$ moderate rewards. From the punishment amount perspective, Q-L and DoubleQ learning schemes outperform SARSA. The best performance in terms of the percentages of TTIs for the maximum rewards is denoted by the ACLA actor-critic policy with a simple parameterization scheme and with a set of discrete actions. To conclude, the CQI aggregation scheme is necessary in the controller state space in order to increase the percentages of TTIs when the system is declared feasible and to minimize at the same time, the amount of punishment rewards in the exploitation stage.

Figure 6.8 shows the performances of the considered policies when the CQI aggregation scheme of $\left(Top3, N_{CT} = 64\right)$ is included in the state space computation. CACLA1 and CACLA2 actor-critic schemes outperform other RL

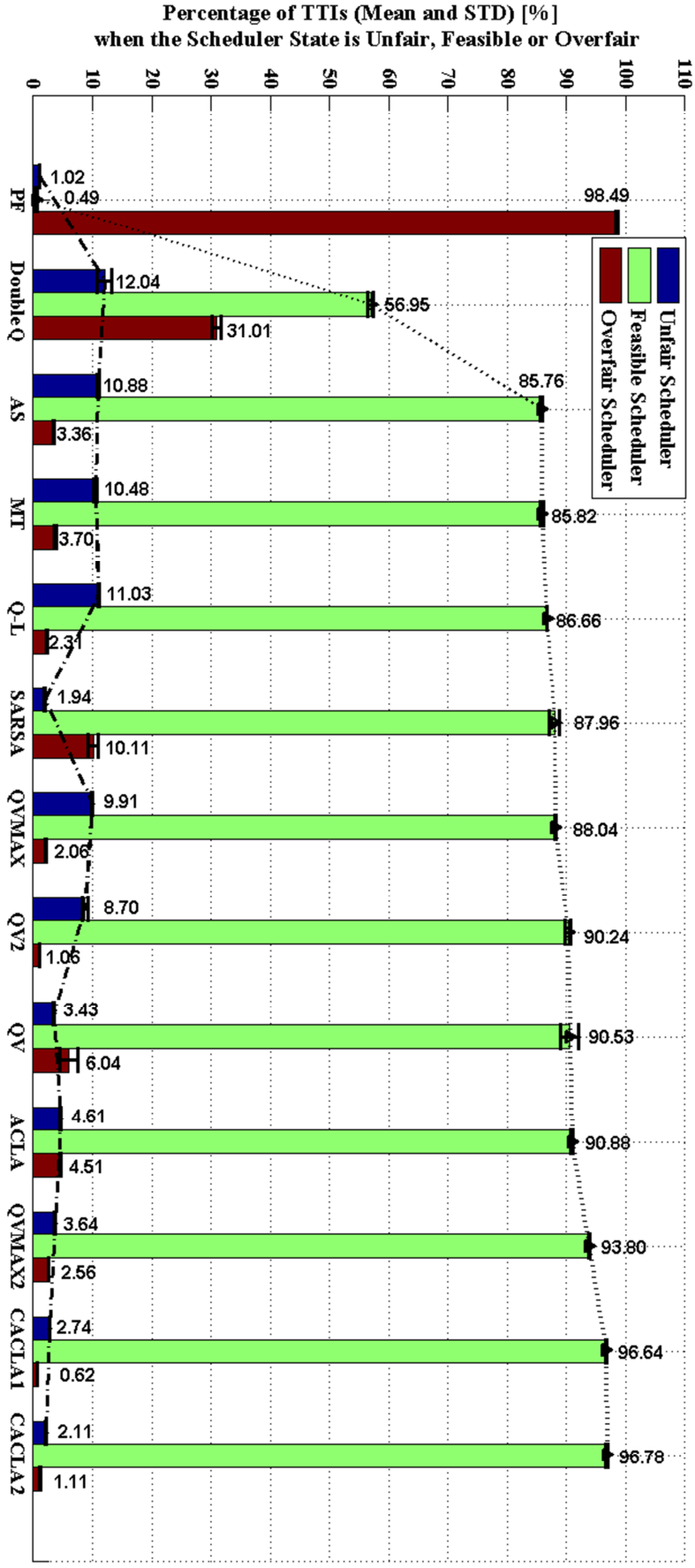


Fig. 6.8 Percentages of TTIs (Mean and Standard Deviation) Under the Unfair, Feasible or Over-fair Scheduler States. CQI Aggregation scheme: $(Top3, N_{CT} = 64)$

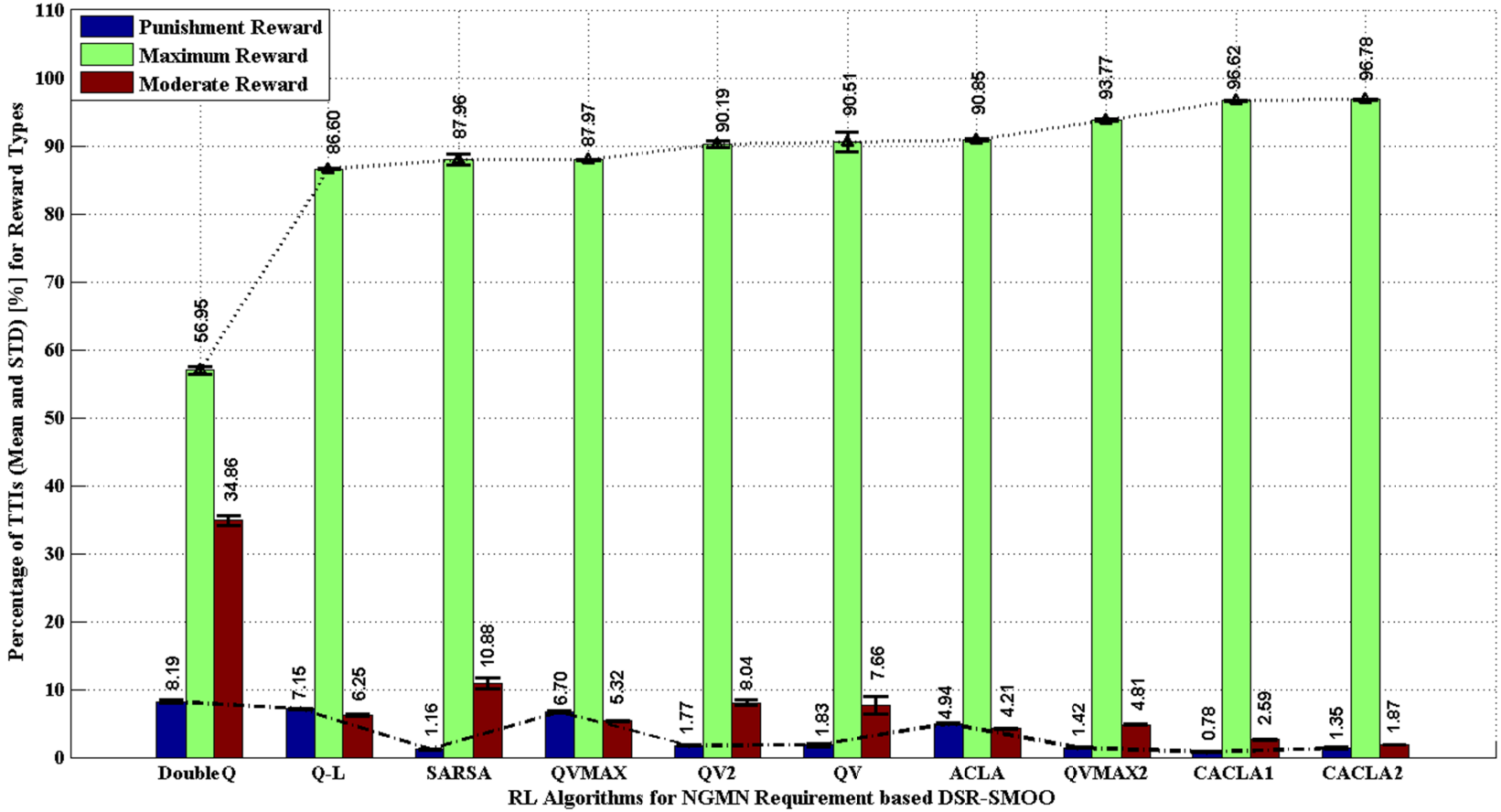


Fig. 6.9 Percentages of TTIs (Mean and Standard Deviation) for the Punishment, Moderate and Maximum Rewards. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

approaches when the controller state is feasible or over-fair. At the same time, the mean percentage of TTIs when the system stays unfair is considerably lower when compared with other methods. Other RL schemes such as ACLA, QVMAX2, QV2 and QV indicate a degradation of the mean percentage of feasible TTIs $\overline{p}_{TTI}^{F,FAF}$ of about 3-8%. It is important to note that CACLA1 or CACLA2 policies are able to increase the number of TTIs when scheduler state is feasible with about 11% and to decrease the number of TTIs with about 7% for the case when the system is declared unfair. As shown in Fig. 6.8, by using a relatively reduced number of preprocessed CQI data centers, the actor-critic schemes receive enough information to detect optimal actions in order to drive the system in the feasibility region and to maintain the desirable state for a longer time than other methods under the fast-fading models and the fluctuating number of bearers.

When the mean percentages of TTIs for different reward types are measured, CACLA1 and CACLA2 are expected to obtain the highest amount of maximum rewards and the minimum percentage of punishments (Fig. 6.9). Basically, this advantage is caused by the continuous action spaces of both actor-critic schemes, whereas the other RL algorithms use predefined steps of fairness parameters such as $\Delta\alpha = \left\{\pm 10^{-4}, \pm 10^{-3}, \pm 10^{-2}, \pm 10^{-1}, \pm 5 \cdot 10^{-2}, 0\right\}$ which are not always optimal in reaching the NGMN fairness feasible state. On the other hand, by using the critic in the state-action updates, the feasible state can be found much faster in the exploration stage. This way, CACLA algorithms are able to learn much faster how to maintain the desired controller state when the network conditions change at each TTI.

The CDF representation of the most relevant RL algorithms is shown in Fig. 6.10 for the AUT-EMF observations when the number of users does not fluctuate (transitions from one state space to another when the number of users is changing are not considered). CACLA2 and CACLA1 algorithms respect the NGMN fairness requirement, showing at the same time a higher system throughput when compared with other approaches such as PF, DoubleQ, QV2, QVMAX and QVMAX2. The static parameterization technique localizes the CDF curve in the over-fairness area leading in fact to the waste in the system capacity.

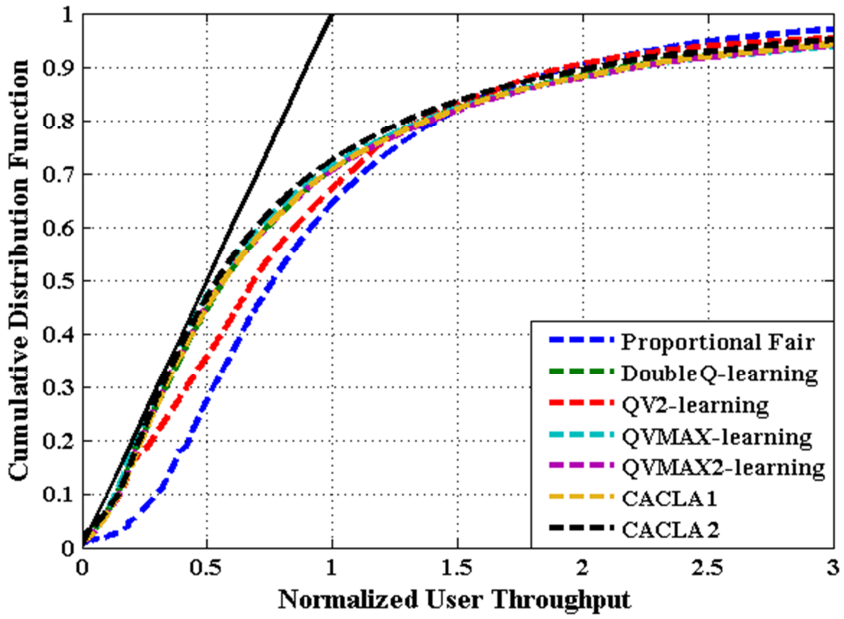


Fig. 6.10 The CDF Curve of the Proposed Policies

The difference between CACLA1 and CACLA2 from the fairness index and system throughput tradeoff point of view is depicted in Fig. 6.11 for the AUT-EMF observations obtained from 10 seconds of the exploitation period. As expected, by performing the GPF-DP technique, CACLA2 is able to increase the system capacity under a fluctuating number of users when compared with the main candidate which is the scheduling policy obtained by CACLA1 actor-critic.

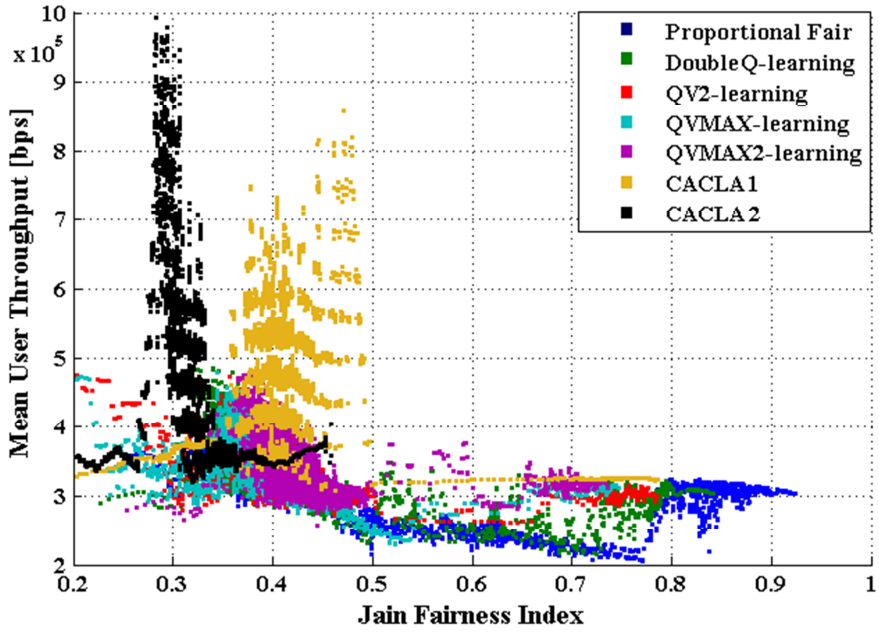


Fig. 6.11 JFI –Mean AUT-EMF Tradeoff

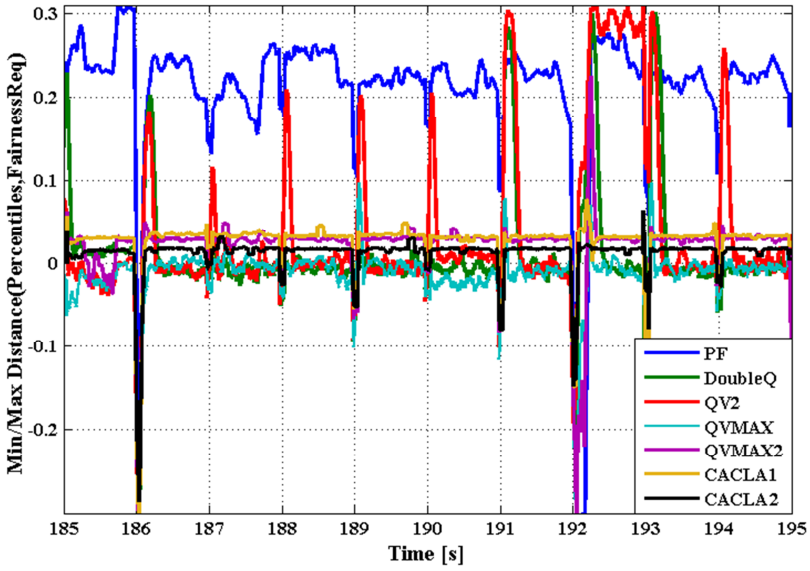


Fig. 6.12 Measured Min/Max Distances from the NGMN Requirement

The measured minimum or maximum distances $\left(d_{max,r}^{CDF,t}, d_{min,r}^{CDF,t}\right)$ from the NGMN requirement can show how fast the learned policy can adapt the newest situations when the traffic load varies drastically. As can be seen from Fig. 6.12, CACLA2 and CACLA1 keep the minimum distance $d_{min,r}^{CDF,t}$ from the fairness region in the required confidence interval of $\xi \in [0, 0.05]$. Other approaches such as QVMAX and DoubleQ learning show an unstable $d_{min,r}^{CDF,t}$ behavior which in fact pushes the system in the unfair region. The QV2 scheduling policy shows a higher amount of moderate rewards, and the scheduler converges in the confidence interval much slower when compared with other techniques.

Figure 6.13 highlights the numerical values of (α_t, β_t) when the system is considered to be feasible. For the GPF-SP parameterization, the optimal range is $\alpha_t \in [0.5; 0.6]$ when the AUT-EMF forgetting factor is $\beta_{\bar{T}} = 0.01$. When $\beta_{\bar{T}} \nearrow$, the optimum range of α_t parameter increases, whereas when $\beta_{\bar{T}} \searrow$, the optimum parameterization range of α_t becomes even lower than $\alpha_t \in [0.5; 0.6]$. It is important to point out that if the forgetting factor is $\beta_{\bar{T}} < 0.001$, then the controller actions are not able to reach the feasible state due to the insignificant contribution of the scheduling procedure in the AUT-EMF computation.

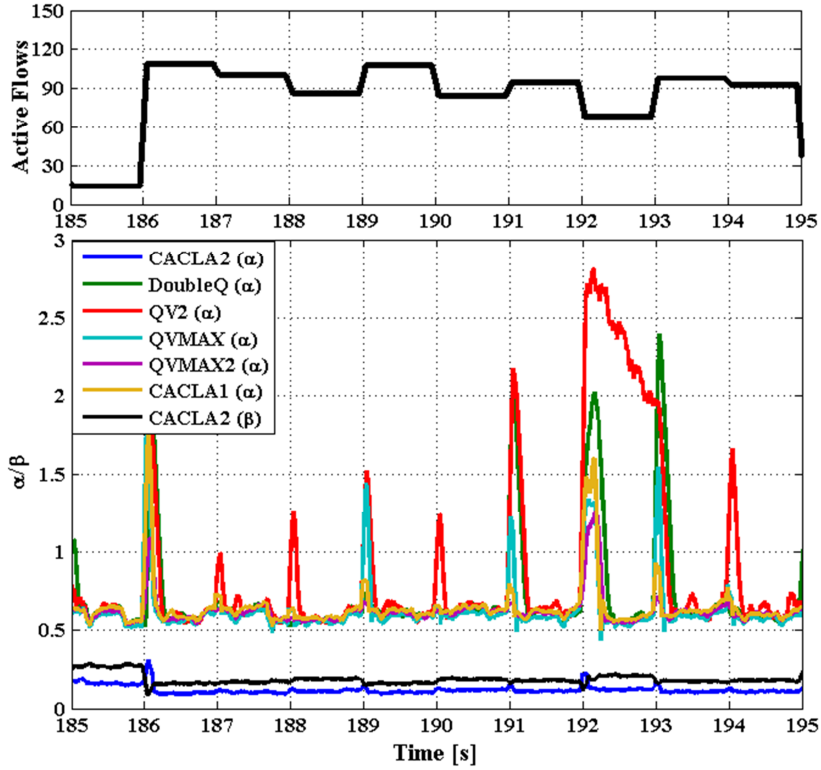


Fig. 6.13 Obtained Parameterization Values

The impact of different CQI aggregation schemes is illustrated in Fig. 6.14 for CACLA2, CACLA1 and MT approaches. The numerical values in terms of $\overline{p}_{TTI}^{F,STAT}$ and $\overline{p}_{TTI}^{F,RW}$ (for both mean and STD values) for other RL algorithms are shown in Appendix F. The mean percentage of TTIs when the scheduler is declared feasible is $\overline{p}_{TTI}^{F,FAF} < 83\%$ for all the analyzed RL approaches when the aggregate channel information is not considered in the controller state space representation. By using the simplest CQI aggregation scheme with $N_{CT} = 64$ number of centers and Top3 mass mode representation, the gain obtained in terms of the mean percentage of feasible TTIs $\overline{p}_{TTI}^{F,FAF}$ is about 13% for both CACLA1 and CACLA2 scheduling policies (Fig. 6.14), outperforming at the same time, the existing approaches (MT and AS) of about 11% of feasible TTIs.

As suggested in Fig 6.14, enhanced performances can be obtained by using the following aggregation schemes: $(Top4, N_{CT} = 256)$, $(Top4, N_{CT} = 512)$ and $(Top5, N_{CT} = 512)$. When a higher number of centers is used, the system

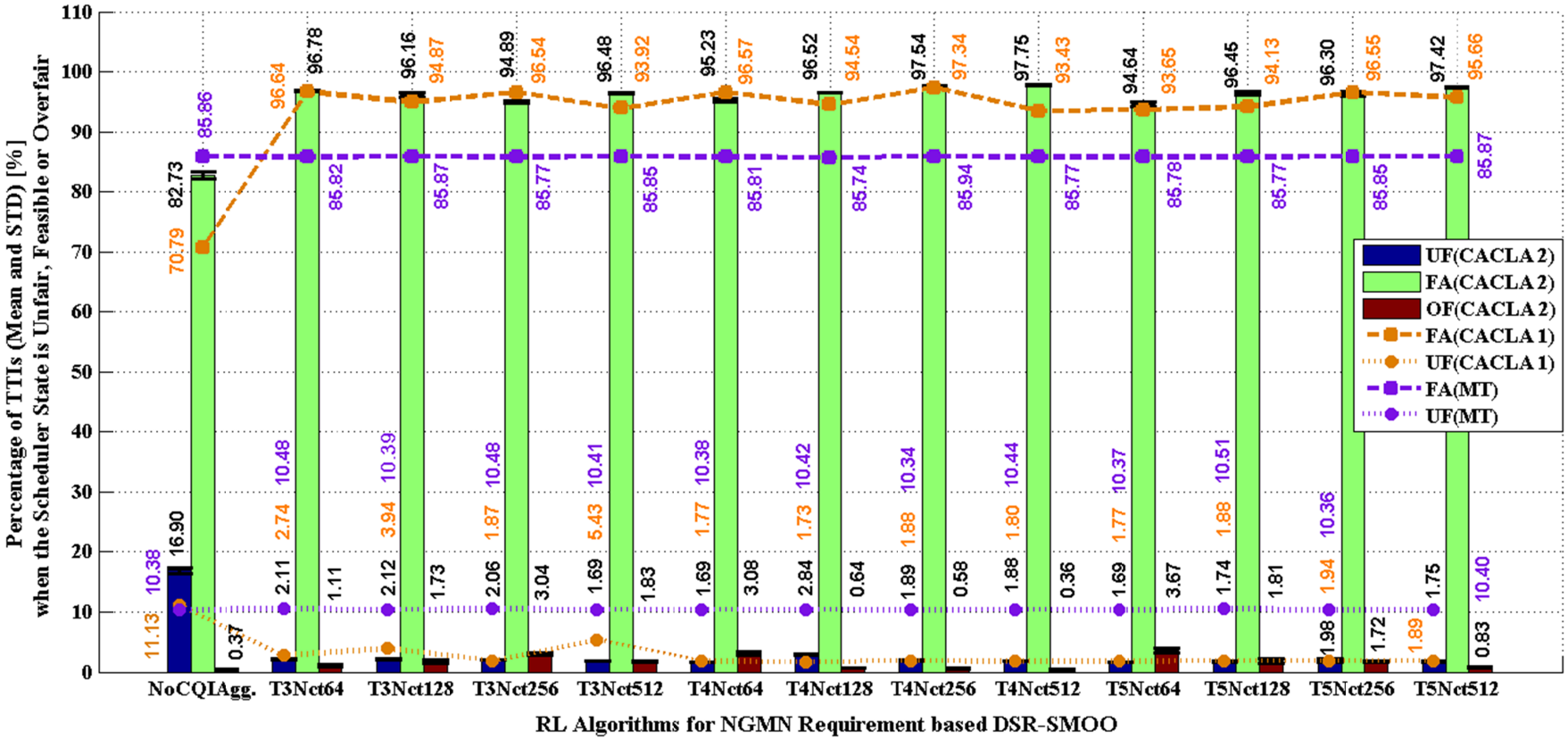


Fig. 6.14 Percentages of TTIs (Mean and Standard Deviation) Under the Unfair, Feasible or Over-fair Scheduler States with Multiple CQI Aggregation Schemes: $\{Top3, Top4, Top5\} : N_{CT} = \{64, 128, 256, 512\}$

complexity increases even when the trained CQI aggregation structure is exploited for different RL stages. For this reason, it is preferable to use the CQI aggregation architecture with $N_{CT} = 64$ number of preprocessed CQI centers even if slightly better precision and scheduling results can be obtained by using a larger number of preprocessed CQI centers. From the percentages of TTIs when the system is unfair, the proposed CQI configurations for CACLA1/CACLA2 actor-critic schemes outperform the MT prediction model of about 8%. The positive impact introduced by the proposed CQI aggregation techniques is clearly shown in Fig. 6.14 in which the quantity of TTIs when the system is declared unfair is reduced by about 14% when compared with the case when the CQI aggregation technique is not used. To conclude, the proposed scheduling policies being oriented on NGMN fairness requirement with AUT-EMF observations are suitable only and only if one of the CQI aggregation schemes is applied in the controller state space.

As mentioned earlier, the extended set of simulation results of the obtained RL scheduling policies being oriented on the NGMN fairness criterion based on the AUT-EMF observations is highlighted in Appendix F. CACLA2 and CACLA1 policies assure the best mean percentage of feasible TTIs, by minimizing at the same time the STD values. The mean percentages of TTIs with punishment and moderate rewards are minimized. From these reasons, the scheduling policies obtained by using the CACLA1 and CACLA2 actor-critic schemes are considered sustainable being able to adapt to the newest conditions when the NGMN fairness criterion is considered for the AUT-EMF observations.

6.2.5.3 DSR-SMOO MDP Based on Average Throughput Observations with Median Moving Filter

The idea of long or short term fairness adaptation is closely correlated with the time window length T_w^E for the AUT-EMF computations and with T_w^M for the user throughput calculated based on the median moving filter. These parameters are considered to be crucial when the intelligent controller is used in order to adapt online the fairness or the GBR satisfaction objectives. For very restrictive lengths of time windows, the entire scheduling results start to fluctuate, the

scheduler rewards are noisy, and the LTE controller is not able to learn the optimal actions. In this case, the short term fairness is addressed. When the time windows are very large (e.g., hundreds of TTIs), the reward value depends on a very large number of observations and the action taken at the current time instant depends also on the history of the controller state transitions. Therefore, the instantaneous reward issued by the LTE scheduler TTI-by-TTI is an accumulated version of many AUT observations for many previous states. This, in fact, makes an impossible job for the controller to quantify the real benefit of action taken in the previous TTI since the reward value obtained in the current state is the subject of observations averaged on the long term purpose.

Another important aspect of the filter window length is represented by the system throughput and user fairness tradeoff concept. When the short term fairness adaptation is performed, the system throughput is seriously degraded based on the considered filter length. When the filter length is large and the long term fairness adaptation is considered, the system throughput can be increased. Therefore, the optimum filter length is required in order to maintain a reasonable system throughput level and to assure accurate reward values in order to help the LTE scheduler controller in taking optimal decisions state-by-state.

For the purpose of this study, the AUT-EMF observations are considered by the marginal utility functions which compute the optimization problem from Eq. 6.5 and the AUT-MMF observations are used to compute the reward functions as a multi-objective evaluator and $\widehat{T}_i[t] = \overline{\overline{T}}_i[t]$, where the AUT-MMF can be computed by using Eq. 6.1. The median filter time window T_w^M depends on the traffic load and on the maximum number of users which can be scheduled at each TTI as shown by Eq. 6.21:

$$T_w^M = \rho \cdot \left\lceil \frac{|\mathcal{U}_t|}{N_{Sched}^{Max}} \right\rceil \tag{6.21}$$

where N_{Sched}^{Max} represents the maximum number of users which can be scheduled at each TTI based on the signaling overhead constraints, and $\rho \in \mathbb{R}^+$ is the ***windowing factor***. The maximum number of schedulable users N_{Sched}^{Max} affects the

HoL delay objective (and more aspects about this parameter are discussed in Chapter 7). The windowing factor can take constant or variable values during the scheduling procedures. When the windowing factor is dynamic $(\rho = \rho_t)$, it has to be decided together with other controller actions. In this sense, CACLA2+ which has the state space of three continuous actions $(\Delta\alpha_t, \Delta\beta_t, \Delta\rho_t)$ is proposed in Chapter 7 in order to adapt the windowing factor during the downlink transmission when other scheduling objectives are considered. In this sub-section, the optimal static windowing factor is determined through extensive simulation results in order to adapt the NGMN fairness on a short-term purpose and to assure a reasonable tradeoff level between user fairness and system throughput.

In order to find the optimum windowing factor which increases the percentage of TTIs when the system is feasible, the same simulation parameters from Tables 6.1 and 6.2 are considered. The only difference is the fact that different MLPNN functions are trained based on different RL algorithms for different windowing factors in the interval of $\rho \in [2.0, 5.5]$ with the factor step of 0.25. It is important to specify that the optimum range for the windowing factor which is proposed in the current sub-section considers the fluctuating number of users in the interval of [15,120] and that the maximum number of users which can be scheduled at each TTI is $N_{Sched}^{Max} = 10$. When the windowing factor is $\rho = 2.0$ and the number of active users is $|\mathcal{U}_t| = 15$, the minimum time window length becomes $T_w^M = 3$, whereas when the factor is $\rho = 5.5$ and the traffic load increases to $|\mathcal{U}_t| = 120$, then the filter length is $T_w^M = 66$.

The impact of the learned policies in the CDF domain is highlighted in Figures 6.15, 6.16 and 6.17. For the very restrictive filter length $(\rho = 2.5)$, the obtained policies are not able to respect the NGMN requirement and to localize the scheduler in the unfair area. When the windowing factor is $\rho = 4.0$ and the filter length belongs to $T_w^M \in [6; 48]$, then CACLA1, CACLA2 and ACLA policies are able to respect the NGMN requirement, whereas other policies such as QV2, QVMAX, QVMAX2, MT and AS localize the scheduler in the unfair region.

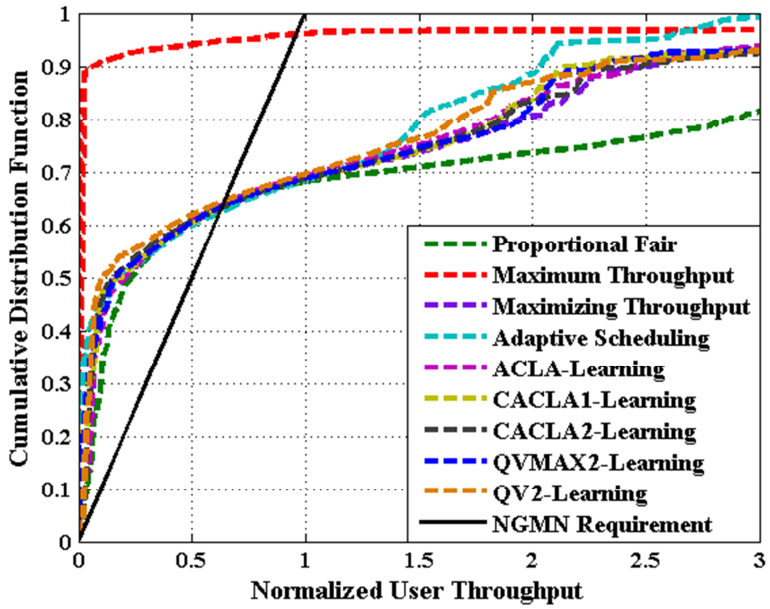


Fig. 6.15 CDF for Static Windowing Factor $(\rho = 2.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$

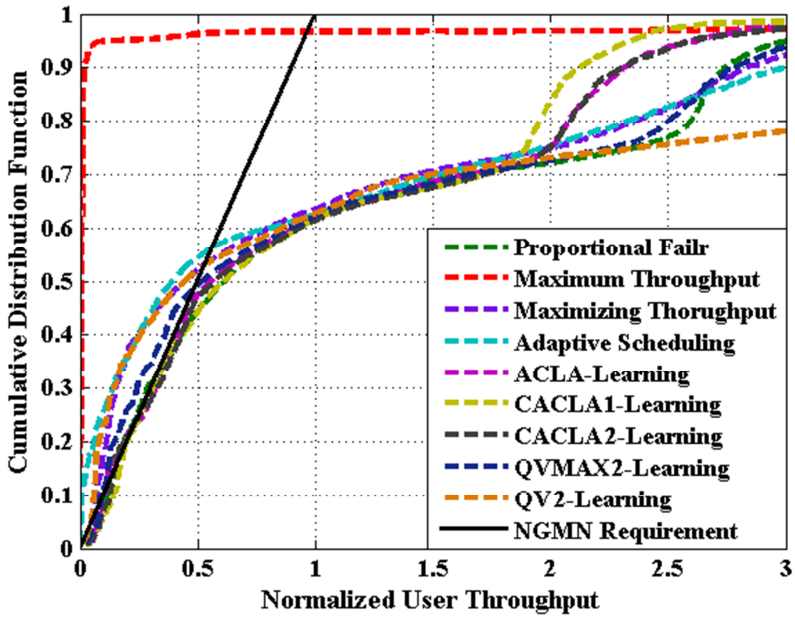


Fig. 6.16 CDF for Static Windowing Factor $(\rho = 4.0)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$

When the filter length increases to $T_w^M \in [8;66]$ for $\rho = 5.5$, the policies start to fluctuate, and CACLA1 and QV2 learning procedures push the scheduler in the over-fair area while the other approaches maintain the unfair region for large time

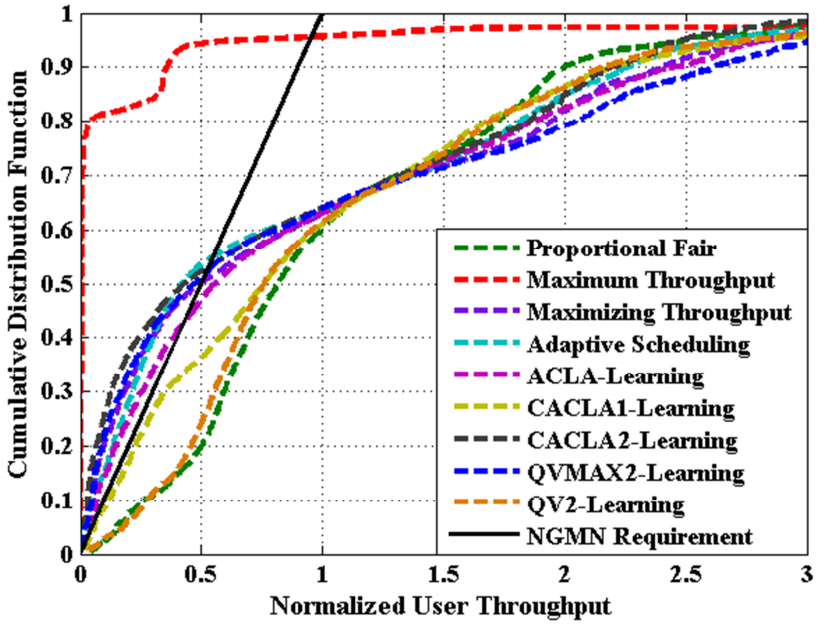


Fig. 6.17 CDF for Static Windowing Factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$

periods of the exploitation procedure. Based on Figures 6.15, 6.16 and 6.17, when the scheduler is feasible, the percentages of TTIs can be increased if the windowing factors belong to the optimal interval of $\rho \in [3.0, 4.0]$.

Figure 6.18 shows the percentages of TTIs when the controller state is $\mathcal{S}_t^{C,F} \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for $\rho = 3.5$ and $\{Top3, N_{CT} = 64\}$ configuration for the CQI aggregation structure. From the perspective of the mean percentages of TTIs when the system is over-fair $\left(\overline{p}_{TTI}^{F,OFF}\right)$, the static parameterizations of PF and MF scheduling rules together with the existing approaches (MT and AS) indicate the worst performances. On the other side, CACLA1 and CACLA2 outperform MT and PF with more than 11% when the mean percentage of TTIs $\overline{p}_{TTI}^{F,UFF}$ is considered. QVMAX2 and ACLA policies achieve a level of $\overline{p}_{TTI}^{F,FAF} = 82\%$ and perform better than CACLA1 and CACLA2 actor-critic schemes. When compared with the main candidates, QXMAX2 gains more than 20% of feasible TTIs when compared with the MT prediction model and more than 5% when compared with the adaptive scheduling method (AS).

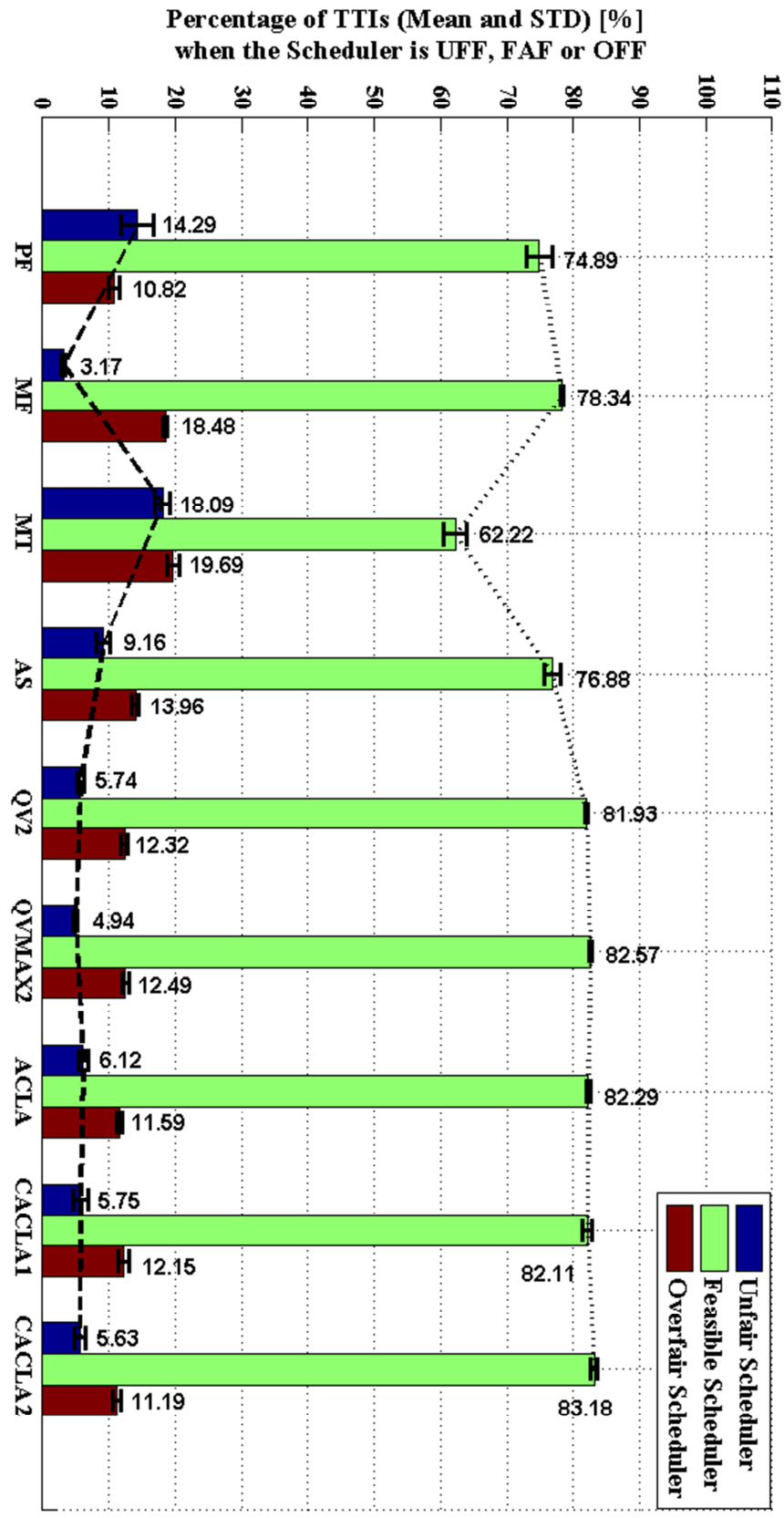


Fig. 6.18 Percentages of TTIs (Mean and Standard Deviation) Under the Unfair, Feasible or Over-fair Scheduler States for the Windowing Factor Value of $\rho = 3.5$ and CQI Aggregation Scheme: $\{Top3, N_{CI} = 64\}$

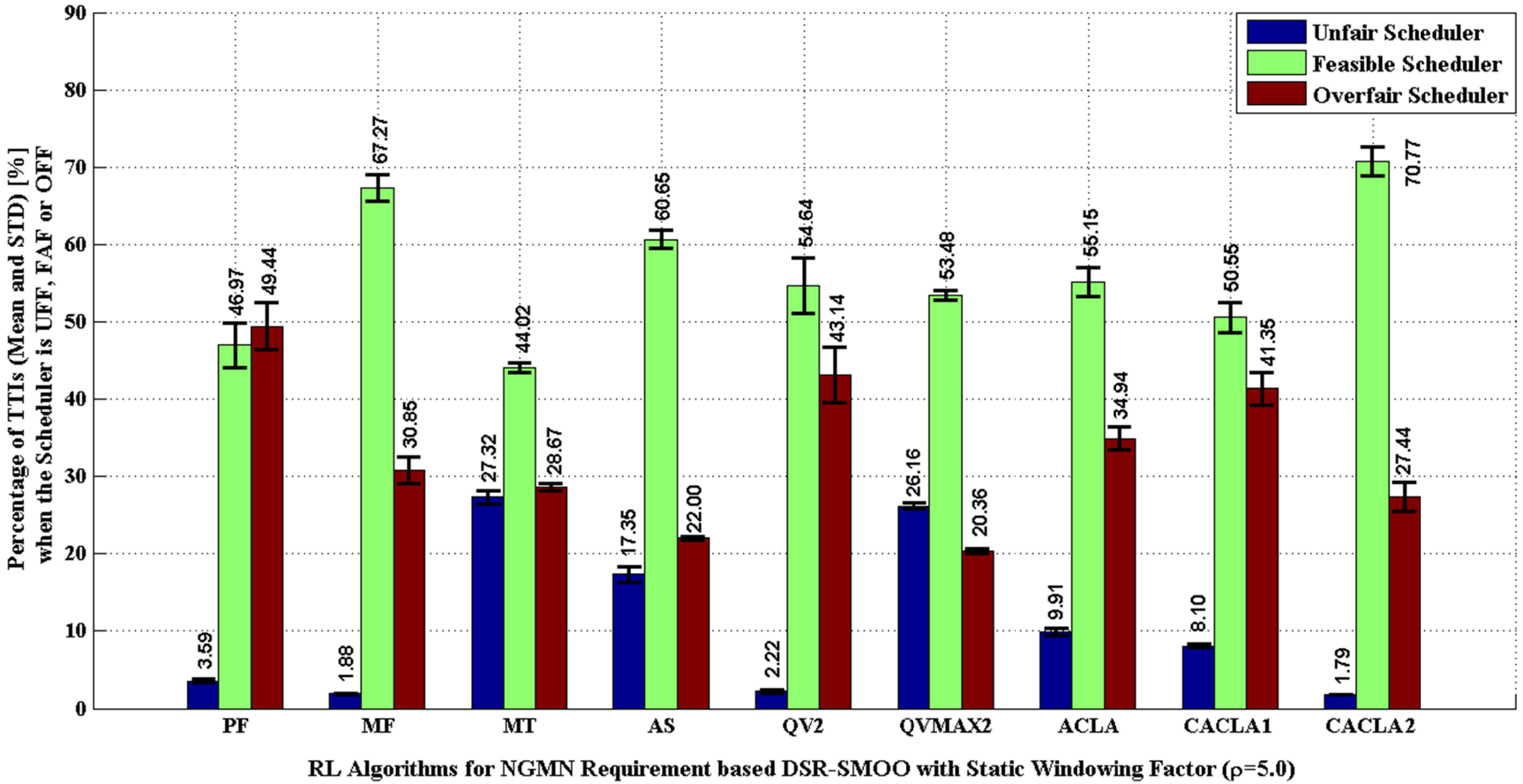


Fig. 6.19 Percentages of TTIs (Mean and Standard Deviation) Under the Unfair, Feasible or Over-fair Scheduler States for the Windowing Factor Value of $\rho = 5.0$ and CQI Aggregation Scheme: $\{Top3, N_{CT} = 64\}$

When the static windowing factor is $\rho = 5.0$ (Fig. 6.19), CACLA2 is the best option from the viewpoints of the percentages of TTIs when the scheduler stays feasible and unfair. Other RL approaches offer relatively very close performances with the amendment that QVMAX2 indicates the lowest percentage of TTIs when the scheduler is over-fair. When compared with the adaptive scheduling (AS) scheme, CACLA2 gains more than 10% of feasible TTIs due to the advantage introduced by the CQI aggregation scheme.

The testing reward performance of the exploited policies is analyzed in Figs. 6.20 and 6.21. When the windowing factor is $\rho = 3.5$ (Fig. 6.20), the RL approaches localize the NGMN feasible region by achieving a mean percentage of maximum rewards of $\bar{p}_{TTI}^{F,MRW} > 80\%$ for every learned policy. When compared against the QVMAX2 policy, CACLA2 prefers to receive more punishments than moderate rewards until the AUT-MMF observations stabilize in the NGMN fairness region. For the same policy, the mean percentage of moderate rewards is reduced with about 4-7% when compared with other candidates. In the case of $\rho = 5.0$, most of the proposed policies are not able to reach the desired state since

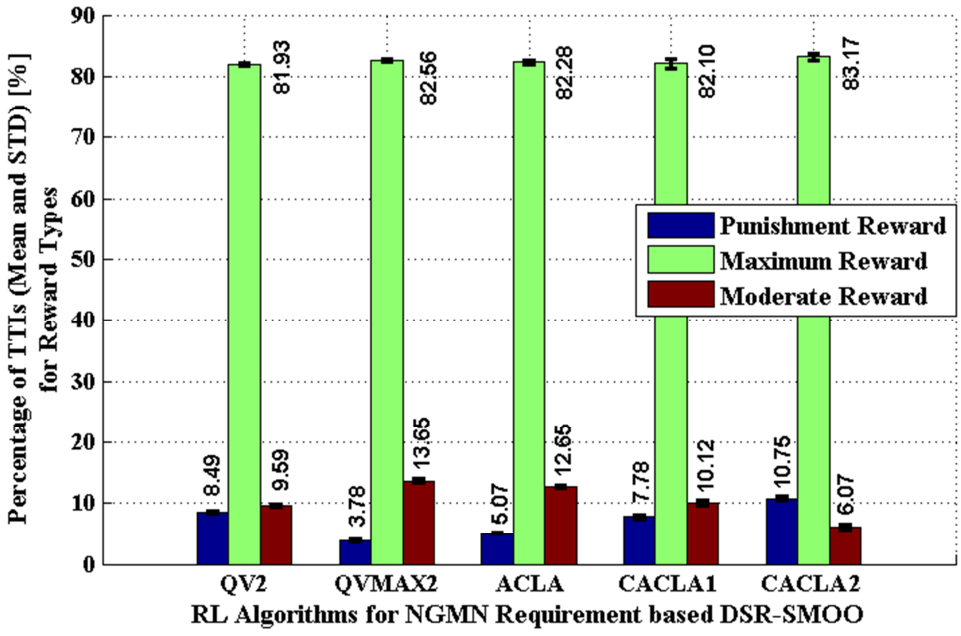


Fig. 6.20 Percentages of TTIs (Mean and Standard Deviation) for the Punishment, Moderate and Maximum Rewards for Windowing Factor $\rho = 3.5$ and CQI Aggregation Scheme: $\{Top3, N_{CT} = 64\}$

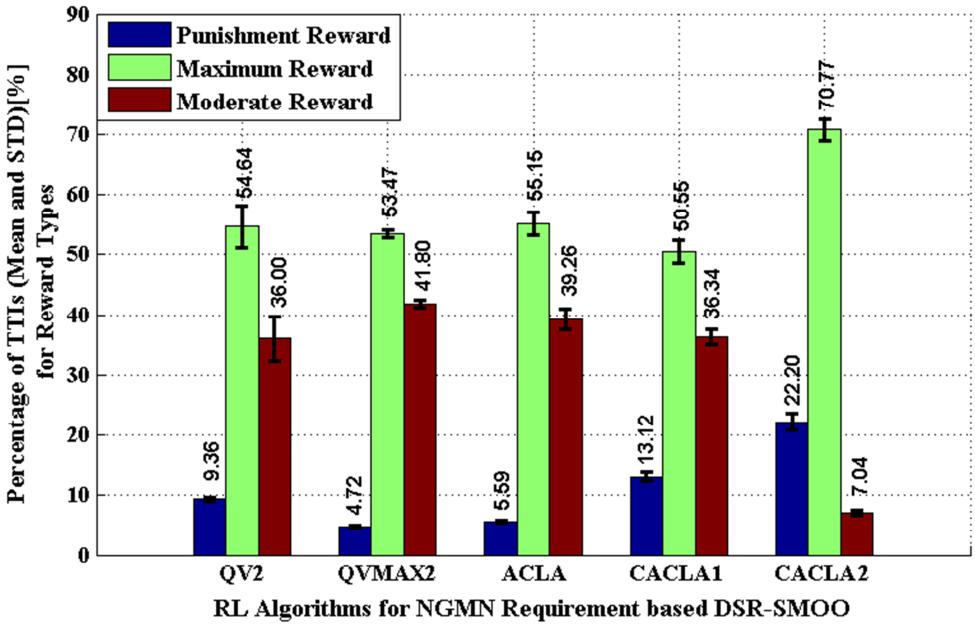


Fig. 6.21 Percentages of TTIs (Mean and Standard Deviation) for the Punishment, Moderate and Maximum Rewards for Windowing Factor $\rho = 5.0$ and CQI Aggregation Scheme: $\{Top3, N_{CT} = 64\}$

the instantaneous scheduler reward is not stabilized enough due to the larger filter length which computes the AUT-MMF observations. For this reason, QV2, QVMAX2, ACLA and CACLA1 spend more time in reaching the optimal state (with more moderate rewards) while CACLA2 waits until the reward value becomes stable while receiving a higher amount of punishment rewards.

Figure 6.22 analyses the percentage of TTIs evolution for different scheduler state status when the windowing factor takes values in the interval of $\rho \in [2.0;5.5]$ by using a factor step of 0.25. The CACLA2 learning scheme is compared against CACLA1 and MT approaches by using the aforementioned windowing factor interval. The highest amount of TTIs when the scheduler is declared feasible, from the NGMN fairness requirement point of view, is obtained when the windowing factor takes the following optimal values $\rho \in [3.0;4.0]$ for the maximum number of schedulable bearers of $N_{Sched}^{Max} = 10$ and a varying traffic load of $|\mathcal{U}_t| \in [15,120]$. When $\rho < 3.0$, the rewards start to fluctuate by pushing the system in the unfair or feasible regions. When $\rho > 4.0$, the impact of the

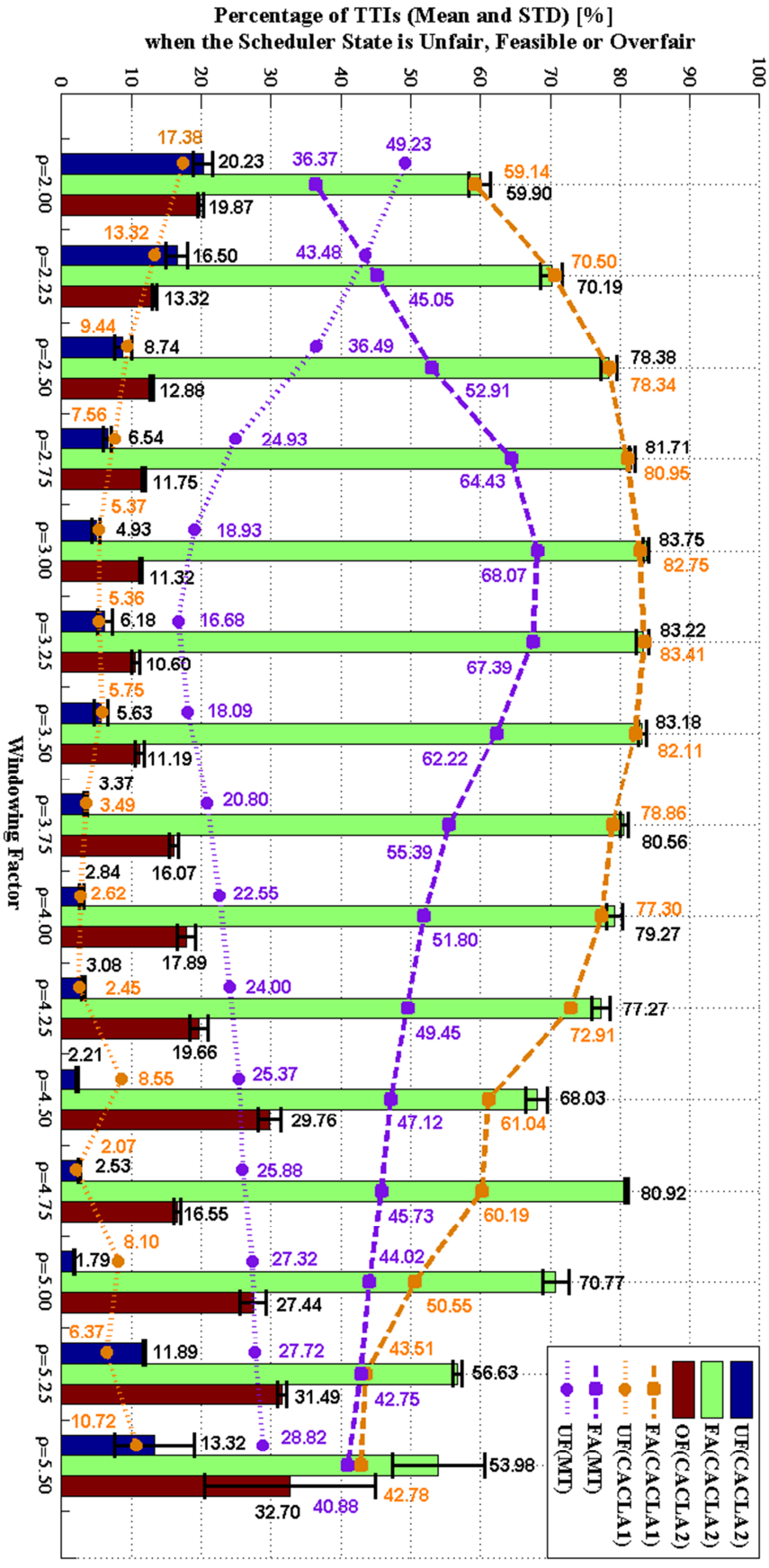


Fig. 6.22 Percentages of TTIs (Mean and Standard Deviation Under the Unfair, Feasible or Over-fair Scheduler States with Variable Windowing Factor of $\rho \in [2.0;5.5]$ and CQI Aggregation Scheme: $\{Top3, N_{CT} = 64\}$

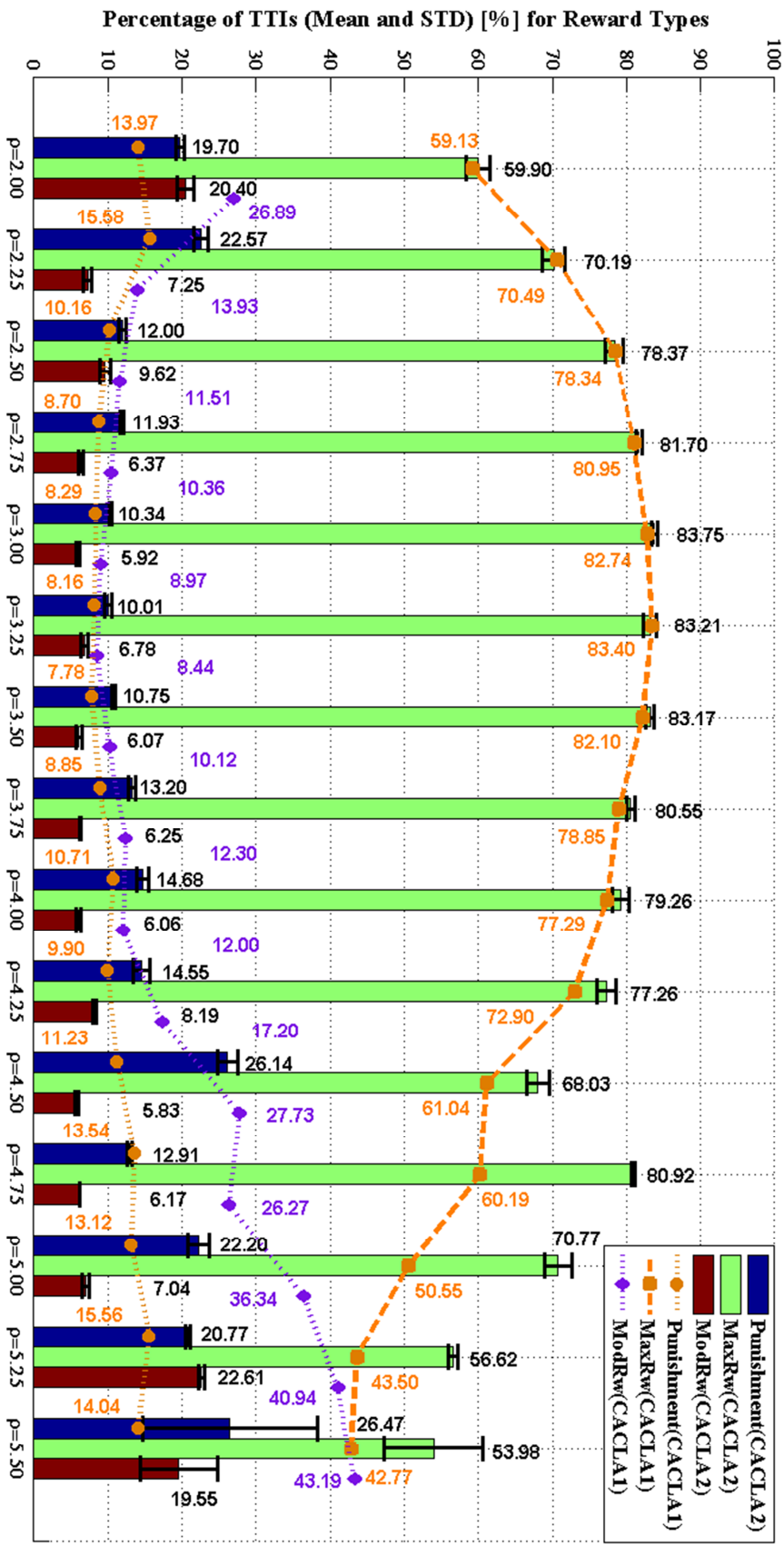


Fig. 6.23 Percentages of TTIs (Mean and Standard Deviation) for Punishment, Moderate and Maximum Rewards with Variable Windowing Factor of $\rho \in [2.0;5.5]$ and CQI Aggregation Scheme: $\{Top3, N_{CT} = 64\}$

continuous action set in the GPF-SP optimization problem is sensed in an accumulated manner, and the number of TTIs, when the scheduler is over-fair, is increased. At the same time, the variation of percentages becomes notable when $\rho \geq 5.0$. CACLA1 offers comparable performance from the $\overline{p}_{TTI}^{F,FAF}$ perspective by imposing a windowing factor of $\rho \leq 4.0$. Beyond this interval, a part of the percentage of TTIs when the scheduler is feasible is taken by $\overline{p}_{TTI}^{F,UFF}$ or $\overline{p}_{TTI}^{F,OFF}$. By comparing the percentages of feasible TTIs, CACLA2 outperforms very clearly the MT prediction methodology by providing a gain higher than 15% for each of the considered static windowing factor. The gain becomes even higher when the percentages of unfair TTIs are taken into account. When $\rho = 2.0$, CACLA2 policy outperforms MT with more than 31% TTIs when the scheduler is declared unfair.

When the given windowing factor interval is considered, CACLA1 provides a lower amount of punishments and a higher quantity of TTIs when the scheduler grants the controller with moderate rewards (Fig. 6.23). As said, by using two continuous actions for the GPF-DP parameterization, the scheduler controller decides that by receiving more punishments when $\rho > 4.0$, the feasible state can be reached much faster. For this reason, it can be concluded that CACLA2 offers better performances when matched against other existing or RL candidates from the viewpoint of NGMN fairness objective with AUT-MMF observations. The numerical results of other RL approaches are presented in Sub-section F.3 from Appendix F for both state status and reward type performances.

From Figures 6.22 and 6.23 it can be concluded that when the windowing factor belongs to $\rho \in [3.0;4.0]$ and the number of active users varies in the interval of $|\mathcal{U}_t| \in [15,120]$, then the scheduling policies trained by using CACLA1 and CACLA2 RL techniques offer very good sustainability by maximizing the percentage of feasible TTIs in the long term purpose under the most severe fluctuations of the radio channel and traffic conditions. This affirmation is valid only if the CQI aggregation scheme is performed. For computational complexity reasons, the simplest CQI aggregation scheme $\{Top3, N_{CT} = 64\}$ is used in the controller state space computation for the rest of the considered simulation results.

6.3 DSR-SMOO MDP Focusing on GBR Objective

The QoS satisfaction differs from the user fairness and system throughput tradeoff objective since the service provided for each active radio bearer should respect a given QoS profile (GBR, HoL delay and PDR requirements). For the GBR SMOO problems, the way of how the user data rate is computed plays a crucial role. In the previous section, two modes of averaging the user throughput have been introduced in terms of AUT-EMF and AUT-MMF. When large filter lengths are used in averaging the user throughput, the probability of satisfying the GBR objectives increases since this performance is measured on the long term purpose by improving the system throughput and degrading the user fairness. But the long term GBR satisfaction does not guarantee the required amount of data in the short term downlink scheduling. Therefore, one purpose of this section is to find the optimal windowing factor which can permit to deliver the requested data rate on a short term purpose.

As seen in Chapter 3, there are three scheduling rules which are focused on the GBR objective. The DSR-SMOO MDP selects at each TTI the best scheduling rule to be applied in order *to increase the number of TTIs when the active bearers are 100% satisfied* from the GBR objective point of view. The controller scheduler state space should contain some additional indicators which are oriented on the GBR objective satisfaction. The action space becomes fully discrete for all RL algorithms, and each controller action represents in fact the index of the scheduling rule to be selected in the current scheduling time instant.

6.3.1 DSR-SMOO Problem Focusing on GBR Objective

The DSR-SMOO problem focusing on the GBR objective includes the MU functions presented in Chapter 3 corresponding to the GPF-BF, GPF-RAD and GPF-mM scheduling rules with the amendment that the weight functions are computed *based on the AUT-MMF observations*. The scheduling rule proposed in this section is entitled GPF based on Lagrange Multiplier (GPF-LM) where the Lagrange multiplier $\lambda_i^G[t]$ is determined according to Eq. 6.22:

$$\lambda_i^G[t] = \lambda_i^G[t-1] + \beta_{\lambda^G} \cdot \left(T_i[t] - \breve{\overline{T}}_i[t] \right) \tag{6.22}$$

where β_{λ^G} is the forgetting factor which corresponds to the Lagrange multiplier computation for the GBR objective. Then, the utility function, the weight function and the scheduling rule which are associated with the GPF-LM discipline are expressed by Eq. 6.23, where $\omega_{4,i}^3$ is the GPF-LM constant value:

$$\begin{cases} U_{4,i}^3\left(\overline{T}_i[t]\right) = log\left(\omega_{4,i}^3 + \lambda_i^G\right) \cdot U_{2(\alpha,\beta),i}^2\left(\overline{T}_i[t]\right) \\ F_{4,i}^{'3}\left(\overline{T}_i[t]\right) = 1 \Big/ \left(\overline{T}_i[t]\right)^{\alpha} \\ W_{4,i}^3\left(T_i[t]\right) = log\left(\omega_{4,i}^3 + \lambda_i^G\right) \cdot \left(r_{i,j}[t]\right)^{\beta-1} \\ D_{4,i}^3\left(T_i[t]\right) = log\left(\omega_{4,i}^3 + \lambda_i^G\right) \cdot \left(r_{i,j}[t]\right)^{\beta} \Big/ \left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \tag{6.23}$$

By unifying the analyzed marginal utility functions oriented on the GBR requirement, the DSR-SMOO problem focusing on GBR objective follows the form exposed in Eq. 6.24.a. The fairness parameters are fixed to $(\alpha = 1, \beta = 1)$ for the entire scheduling session. The decisions $\{c_{3,1}[t], c_{3,2}[t], c_{3,3}[t], c_{3,4}[t]\}$ represent the controller action index mapped in the scheduling rule decision for the optimization problem. For instance, if $c_{3,1}[t] = 1$, the selected scheduling rule

$$\begin{aligned} (P_G): \max_{\pi_{RB}[t]} \Bigg\{ & c_{3,1}[t] \cdot \left[\sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} u_{1,i}^3[t] \cdot b_{i,j}[t] \cdot \left(1 + \omega_{1,1}^3 \cdot e^{-\omega_{2,1}^3 \cdot \left(\overline{\overline{T}}_i[t] - \breve{\overline{T}}_i[t] \right)} \right) \cdot \frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}} \right] \; 1: \begin{pmatrix} GPF \\ BF \end{pmatrix} \\ & + c_{3,2}[t] \cdot \left[\sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} u_{2,i}^3[t] \cdot b_{i,j}[t] \cdot e^{\omega_{2,i}^3 \cdot TC_i[t]} \cdot \frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}} \right] \; 2: (GPF - mM) \\ & + c_{3,3}[t] \cdot \left[\sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} u_{3,i}^3[t] \cdot b_{i,j}[t] \cdot \frac{\breve{\overline{T}}_i[t]}{\overline{\overline{T}}_i[t]} \cdot \frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}} \right] \; 3: (GPF - RAD) \\ & + c_{3,4}[t] \cdot \left[\sum_{i=1}^{|\mathcal{U}_t|} \sum_{j=1}^{|\mathcal{B}|} u_{4,i}^3[t] \cdot b_{i,j}[t] \cdot log\left(\omega_{4,i}^3 + \lambda_i^G\right) \cdot \frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}} \right] \Bigg\} \; 4: \begin{pmatrix} GPF \\ LM \end{pmatrix} \end{aligned} \tag{6.24.a}$$

$$
\begin{aligned}
& c_{3,1}[t]+c_{3,2}[t]+c_{3,3}[t]+c_{3,4}[t]=1 \\
& \sum_{w_3=1}^{4} u^3_{w_3,i}[t]=1,\ i=1,...,|\mathcal{U}_t| \\
& \sum_{i=1}^{|\mathcal{U}_t|} u^3_{w_3^*,i}[t]=|\mathcal{U}_t|,\ w_3^* \in \mathcal{PU}_3 \\
(C_G)\ s.t.: & \sum_{i=1}^{|\mathcal{U}_t|} u^3_{w_3^{\otimes},i}[t]=0,\ w_3^{\otimes}=1,...,|\mathcal{PU}_3|, \forall w_3^{\otimes} \neq w_3^* \\
& \sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t]=1,\ j=1,...,|\mathcal{B}| \\
& b_{i,j}[t] \in \{0,1\},\ \forall i \in \mathcal{U}_t, \forall j \in \mathcal{B} \\
& \{u^3_{1,i}[t], u^3_{2,i}[t], u^3_{3,i}[t], u^3_{4,i}[t]\} \in \{0,1\}, \forall i \in \mathcal{U}_t \\
& \{c_{3,1}[t], c_{3,2}[t], c_{3,3}[t], c_{3,4}[t]\} \in \{0,1\}
\end{aligned}
\tag{6.24.a}
$$

is the GPF-BF rule [54] introduced in Eq. 3.41, or if $c_{3,4}[t]=1$, the proposed GPF-LM rule is applied at TTI *t*. The same MU function is assigned for each user $i \in \mathcal{U}_t$ at each TTI *t* by using the assignation vectors $\{u^3_{1,i}[t], u^3_{2,i}[t], u^3_{3,i}[t], u^3_{4,i}[t]\}$.

Based on Eq. 6.24, the role of the scheduler controller is to find optimal actions $\mathcal{A}_t^{a,G}, a=\{1(c_{3,1}[t]=1); 2(c_{3,2}[t]=1); 3(c_{3,3}[t]=1); 4(c_{3,4}[t]=1)\}$ in order to maximize the problem (P_G) by respecting the set of constraints (C_G) and by satisfying the objective conditions (O_G) TTI-by-TTI as suggested in Eq. 6.24.b:

$$
(O_G): \qquad \overline{\overline{T}}_i[t] \geq \breve{T}_i[t], \forall i \in \mathcal{U}_t \tag{6.24.b}
$$

The idea of the DSR-SMOO problems being focused on the GBR requirement is to increase the percentage of TTIs when the active bearers are satisfied from the viewpoint of GBR objective. At the same time, the number of punishment rewards should be minimized. When all active bearers are satisfied from the GBR constraint objective (O_G), the feasible or the optimal state is reached and the DSR-SMOO MDP problem becomes episodic. This is why the optimization problem (P_G) correlated with the controller action decision is focused in satisfying all the active bearers at each TTI rather than increasing the satisfaction of some bearers in the detriment of others active users.

6.3.2 Controller State Space for DSR-SMOO Focusing on GBR Objective

The feasibility or the unfeasibility of the controller states for the GBR objectives can be decided strictly based on the intrinsic objectives from Eq. 6.24. Let us define the controller state space for GBR objective such as $\mathcal{S}_t^{C,G}$. The controller state is feasible $\mathcal{S}_t^{C,G} \in \mathcal{FAG}$ when every active bearer satisfies the objective condition (O_G), and the scheduler becomes unfeasible $\mathcal{S}_t^{C,G} \in \mathcal{UFG}$, when there exists at least one bearer which does not receive the requested data rate for a given, let us say, optimum windowing factor. Mathematically, the above concept is expressed in Eq. 6.25:

$$\mathcal{S}_t^{C,G} = \begin{cases} \{\mathcal{FAG}\}, \ if \ \overline{\overline{T}}_i[t] \geq \underline{T}_i[t], \forall i \in \mathcal{U}_t \\ \{\mathcal{UFG}\}, \ if \ \exists \overline{\overline{T}}_i[t] < \underline{T}_i[t] \end{cases} \tag{6.25}$$

The controller state space $\mathcal{S}_t^{C,G}$ considers the elements obtained through the CQI aggregation schemes and the additional observations for arrival rates and queue sizes when other traffic models are considered such as CBR or VBR. The controller state space representation for the GBR DSR-SMOO problems is defined by Eq. 6.26:

$$\begin{aligned} \mathcal{S}_t^{C,G} = \Big\{ & \mathcal{A}_{t-1}^{a,G}, N_{CL}^{A}, \sigma_{CL}^{t}, ks_t, d_{ks}^{t}, \mu_{\overline{T}}^{t}, \sigma_{\overline{T}}^{t}, \mu_{\overline{\overline{T}}}^{t}, \sigma_{\overline{\overline{T}}}^{t}, \mu_{\lambda_{\overline{T}}^{G}}^{t}, \sigma_{\lambda_{\overline{\overline{T}}}^{G}}^{t}, \\ & \mu_{qTX}^{t}, \sigma_{qTX}^{t}, \mu_{\overline{\lambda}}^{t}, \sigma_{\overline{\lambda}}^{t}, N_{t,G}^{SAT}, N_{t,G}^{UNSAT}, |\mathcal{U}_t|, \underline{T}_t \Big\} \end{aligned} \tag{6.26}$$

where $\mathcal{A}_{t-1}^{a,G}$ is the controller action applied in the previous TTI, the instantaneous GBR Lagrange multiplier being defined as $\lambda_{\overline{\overline{T}}}^{G}[t] = \left[\overline{\overline{T}}_i[t] - \underline{T}_i[t] \right]_+$ determines the difference between the AUT-MMF observation and its GBR requirement at each TTI and for each active user $i \in \mathcal{U}_t$, q^{TX} is the queue size for a given active bearer and $\overline{\lambda}_i$ is the average arrival rate. For all of these parameters, the mean and STD values are determined based on Equations 4.6 and 4.7 from Chapter 4.

Additionally, $N_{t,G}^{SAT}$ represents the normalized number of satisfied bearers from the viewpoint of the GBR objective and $N_{t,G}^{UNSAT} = 1 - N_{t,G}^{SAT}$ is the number of unsatisfied bearers at TTI *t*. When $N_{t,G}^{UNSAT} = 0$, then the scheduler is declared optimal from the GBR perspective and the controller state becomes $\mathcal{S}_t^{C,G} \in \mathcal{FAG}$, and otherwise, the controller state is unfeasible and $\mathcal{S}_t^{C,G} \in \mathcal{UFG}$.

6.3.3 Reward Function for DSR-SMOO Focusing on GBR Objective

The scheduler reward function takes into account the reward values obtained for each active bearer. The reward function for each user who requests a predefined level of data rate can be calculated as a normalized value of the difference between the AUT-MMF observation and its GBR requirement as expressed by the following equation:

$$\hat{\lambda}_{i,\overline{\overline{T}}}^{G}[t] = \frac{\overline{\overline{T}}_i[t] - \breve{T}_i[t]}{\breve{T}_i[t]}, \ \forall i \in \mathcal{U}_t \tag{6.27}$$

When $\hat{\lambda}_{i,\overline{\overline{T}}}^{G}[t] > 0$, the bearer $i \in \mathcal{U}_t$ is considered to be satisfied from the viewpoint of the GBR constraint. For MLPNN convergence reasons, the reward function for each active bearer is modeled by using Eq. 6.28:

$$\mathcal{RW}_i^G[t] = \begin{cases} \hat{\lambda}_{i,\overline{\overline{T}}}^{G}[t], \ if \ \ \hat{\lambda}_{i,\overline{\overline{T}}}^{G}[t] < 0 \\ 1, \ if \ \ \hat{\lambda}_{i,\overline{\overline{T}}}^{G}[t] > 0 \end{cases} \tag{6.28}$$

where the definition domain is $\mathcal{RW}_i^G : \mathbb{R} \to \mathbb{R}_{[-1,1]}$. The intrinsic scheduler reward value is obtained by summing the rewards from Eq. 6.28 at each TTI such that:

$$\mathcal{RWI}^G[t] = \sum_{i=1}^{|\mathcal{U}|} \mathcal{RW}_i^G[t] \tag{6.29}$$

The global reward function focusing on the GBR objective should verify if there is any improvement of the total intrinsic reward $\mathcal{RWI}^G[t]$ at each TTI *t* when

compared with $\mathcal{RWI}^G[t-1]$ from the previous TTI in terms of the temporal differences such that:

$$\mathcal{RW}^G[t]=\begin{cases}1, & \text{if } \mathcal{RWI}^G[t]=1\\ \mathcal{RWI}^G[t]-\ell_G\cdot\mathcal{RWI}^G[t-1], & \text{otherwise}\end{cases} \tag{6.30}$$

where the definition domain is $\mathcal{RW}^G:\mathbb{R}^{|\mathcal{U}_t|}\rightarrow\mathbb{R}_{[-1,1]}$ and the parameter $\ell_G=1$ decides that the intrinsic reward from the previous state is taken into account. When the reward is $\mathcal{RW}_t^G=1$, then the controller state is feasible $\mathcal{S}_t^{C,G}\in\mathcal{FAG}$ and when the scheduler reward $\mathcal{RW}_t^G$ is moderate or punishes the controller actions, then the controller state becomes unfeasible and $\mathcal{S}_t^{C,G}\in\mathcal{UFG}$.

6.3.4 Performance Evaluation of Sustainable Scheduling Policies Focusing on GBR Objective

The performance evaluation of the proposed scheduling policies being focused on the GBR objective is performed through two directions: the percentage of TTIs when all active users are satisfied from the viewpoint of GBR requirement and the percentage of TTIs when the scheduler reward is punishment, moderate or maximized. The scheduling policies are learned based on multiple windowing factor settings in order to detect the optimal filter length in the AUT-MMF computation that maximizes the percentage of TTIs when the active users are 100% satisfied from the viewpoint of the GBR objective.

6.3.4.1 Simulation Scenario

The scheduling policies focusing on the GBR objective which use the scheduling rules from the optimization problem (P_G) are trained based on three traffic types: infinite buffer, CBR and VBR. During the exploration and exploitation stages, the traffic load is changed at each 1000 TTIs in the domain of $|\mathcal{U}_t|\in[15;120]$ number or active users and the GBR requirements for the considered traffic types are switched for each active user randomly at each 1000

Table 6.3 LTE Scheduler Parameters for DSR-SMOO Focusing on GBR Objective

Parameters Name	**Description/Values**
System Bandwidth/Cell Radius	20 MHz/1000m [36]
User Speed/Mobility Model	120kmph/Random Direction
Channel Model	Jakes Model (Appendix B)
Path Loss / Penetration Loss	Macro Cell Model / 10 dB [36]
Interfered Cells/ Shadowing STD	0/8dB [36]
Carrier Frequency/DL Power	2GHz/43dBm [36]
Frame Structure	FDD
CQI Reporting Mode	Full-band, periodic at each TTI
PUCCH Model	Errorless
Scheduler Type	GPF-BF[54],[99]/GPF-RAD[101]/ GPF-mM [82],[83]/GPF-LM
$\left\{\omega_{1,1}^{3};\omega_{2,1}^{3};\omega_{2,i}^{3};\omega_{4,i}^{3}\right\}$	$\left\{1.25[99];13.1\cdot10^{-5}[99];10.1[82];2\right\}$
Traffic Type	Infinite Buffer Constant Bit Rate Variable Bit Rate
Max. Number of schedulable users $\left(N_{Sched}^{Max}\right)$ at each TTI	10 (Optimum)
RLC ARQ	Acknowledged Mode (Maximum 5 retransmissions)
AMC Levels	QPSK (1/3, 1/2, 2/3) [36] 16-QAM (1/2, 2/3, 5/6) [36] 64-QAM (2/3, 5/6) [36]
Target BLER	10% (Appendix B)
Number of Users $\left(\lvert\mathcal{U}\rvert\right)$	Variable: 15-120
RL Algorithms	QV, QV2, QVMAX, QVMAX2, ACLA
Discrete MLPNN Actions	$\mathcal{A}_t^G=\begin{cases}1:(GPF-BF)\rightarrow c_{3,1}[t]=1\\2:(GPF-RAD)\rightarrow c_{3,2}[t]=1\\3:(GPF-mM)\rightarrow c_{3,3}[t]=1\\4:(GPF-LM)\rightarrow c_{3,4}[t]=1\end{cases}$
Controller Timescale	1 TTI
Number of MLPNN layers/ Activation Functions	3/ input layer: linear activation, hidden layer: tangent hyperbolic activation, output layer: linear activation
Number of Hidden Nodes	100 (Optimum)
Exploration/Exploitation Periods	500s/95s (Optimum)
AUT-MMF Windowing Factor (ρ)	$\{2.0;2.5;3.0;3.5;4.0;4.5;5.0;5.5\}$

Dynamic GBR Constraints	$\overline{\overline{T}} = \{32;64;128;256;512;1024\}\, kbps$
Maximum HoL Delay ($d_i^{\overline{HoL}}$)	300ms (maximum HoL delay requirement in LTE, Table 2.1)
CQI Aggregation Schemes	Top CQI Mass Modes $\{Top3\} : N_{CT} = \{64\}$
CBR Traffic Type	Data Rates based on GBR Constraints $\lambda = \{32;64;128;256;512;1024\}\, kbps$
VBR Traffic Type	Packet size: Pareto Distrib. $(x = 35.5; \alpha = 1.1)$ [13] Arv. Rate: Geometric Distrib. $(\mu = 1.5; \sigma = 1.93)$ [13]

Table 6.4 LTE Scheduler Controller Parameters for DSR-SMOO Focusing on GBR Objective

RL Algorithm (GBR SMOO)	Learning Rates for Action Values (η^Q)	Learning Rates for State Values (η^V)	Discount Factor (γ)	Exploration Type (ε, τ)
QV	0.001	0.00001	0.99	Boltzmann $(\tau = 10)$
QV2	0.001	0.00001	0.95	Boltzmann $(\tau = 10)$
QVMAX	0.001	0.00001	0.99	Boltzmann $(\tau = 10)$
QVMAX2	0.001	0.00001	0.95	Boltzmann $(\tau = 10)$
ACLA	$\eta_1^Q, \eta_2^Q = 0.0001$	0.0001	0.99	Greedy $(\varepsilon = 5 \cdot 10^{-4})$

TTIs from the set of rate requirements $\overline{\overline{T}}_i[t] \in \{32;64;128;256;512;1024\}$ kbps which is considered the best known GBR requirement set (Table 6.3). From the computational complexity reasons, the maximum number of active users is 120. Those data packets which are waiting in the queue for more than $d_i^{\overline{HoL}} = 300ms$ are dropped and implicitly are declared lost. All users are moving in the macro-cell scenarios with 120kmph by using the random direction mobility model for the exploration and exploitation stages in order to experience as many CQI observations as possible. The arrival rates for the CBR traffic type are chosen according to the GBR requirements and the VBR model is imported from [13]. Based on the extensive simulation results, it is decided to use an optimum number of maximum schedulable users at each TTI of about $N_{Sched}^{Max} = 10$ in the AUT-MMF computations. Each erroneous packet is retransmitted five times and then, it is declared lost and implicitly is dropped from the MAC data queue.

The optimum set of scheduling policies is learned through QV, QV2, QVMAX, QVMAX2 and ACLA algorithms by using static windowing factors in the range of $\rho \in [2.0;5.5]$ with a static factor step of 0.5 when computing the scheduler reward function from Eq. 6.30. When compared with the NGMN fairness objective, the considered RL algorithms require more improvements than evaluations in exploring the scheduling policies (Table 6.4). In this sense, the number of MLPNN nodes is increased; the RL learning rates are decreased for a better policy refinement; and the exploration probability thresholds become $\tau = 10$ and $\varepsilon = 5 \cdot 10^{-4}$ for a better policy improvement procedure during the exploration stage. The discount factors are determined based on the number of feasible TTIs. Based on multiple simulation results, the number of hidden nodes for the MLPNN is set to 100 with the hyperbolic activation function of each hidden node. The entire set of policies is trained by using 500s and then exploited for another 95s. The ER stage is not applicable for this objective.

If $\overline{p}_{TTI}^{G,x}$ represents the mean percentages of TTIs when the percentage of $x\%$ active bearers are satisfied, then the role of the proposed sequential optimization problem is to increase the mean percentage of TTIs $\overline{p}_{TTI}^{G,100\%}$ when the scheduler is feasible averaged over 10 simulations when the learned policy is exploited. For some network conditions, one scheduling rule performs better than any of others, and this way $\overline{p}_{TTI}^{G,100\%}$ can be considerably increased. When the controller state is unfeasible $\mathcal{S}_t^{C,G} \in \mathcal{UFG}$, the role of the learned policies is to increase the mean percentage of TTIs with GBR moderate rewards $\overline{p}_{TTI}^{G,mRW}$ in the detriment of the mean percentage of punishment rewards $\overline{p}_{TTI}^{G,PSH}$. In the exploitation period, the scheduling rules and the learned policies are compared by using the same radio conditions (e.g., multi-path loss, shadowing, interferences) and the same MAC layer conditions (e.g., traffic load, GBR constraints). The results provided in this section represent the general performance when the DSR-SMOO problems focusing on the GBR objective is performed for the downlink scheduling, denoting at the same time the sustainability of the obtained policies.

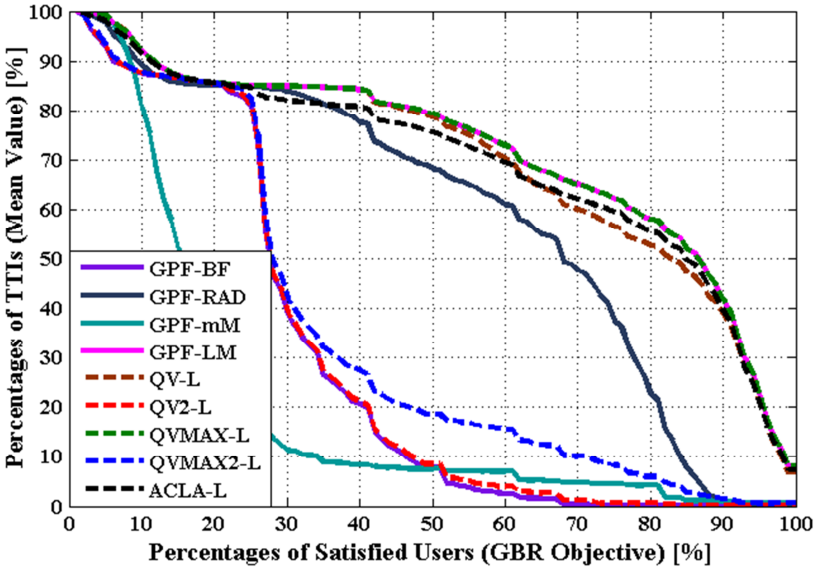


Fig. 6.24 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the Full Buffer Traffic Type and the Windowing Factor of $\rho = 2.5$

6.3.4.2 DSR-SMOO GBR with Full Buffer Traffic

The mean percentages of TTIs for different GBR satisfaction levels $\overline{p}_{TTI}^{G,x}$, when the full buffer traffic type is considered, are highlighted in Figures 6.24, 6.25 and 6.26. As mentioned before, the impact of the windowing factor plays an important role in satisfying the active bearers for a larger number of TTIs when compared with the static scheduling rules. If the windowing factor is $\rho = 2.5$, then the scheduler reward is very noisy and the best option is the proposed GPF-LM scheduling rule. QVMAX, QV and ACLA policies offer the best performances when compared with other RL approaches due to the fact that these policies follow the proposed static policy of the GPF-LM discipline. By using the GPF-LM, QVMAX, QV or ACLA approaches, the gain in percentage of TTIs, when all the bearers are satisfied regardless the network conditions, traffic load or GBR constraint, is of about 8% when compared with other methods. When the static factor of $\rho = 4.0$ is considered, ACLA, QVMAX, and GPF-LM policies outperform other candidates by about 22% from the viewpoint of $\overline{p}_{TTI}^{G,100\%}$. The gain of the mean percentage $\overline{p}_{TTI}^{G,100\%}$ is even higher when the windowing factor

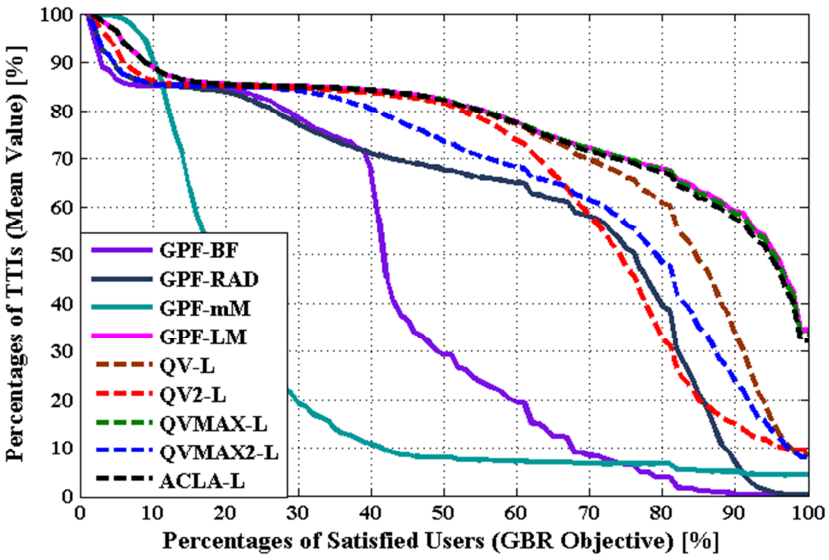


Fig. 6.25 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the Full Buffer Traffic Type and the Windowing Factor of $\rho = 4.0$

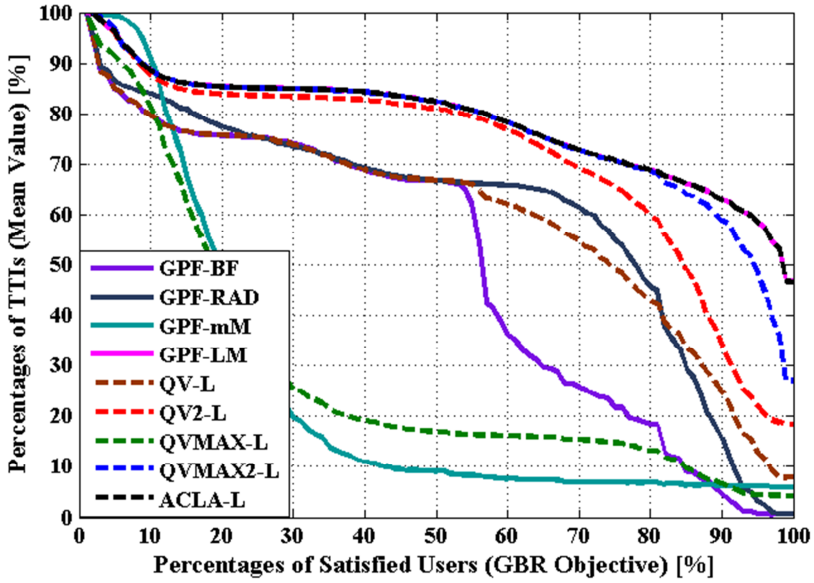


Fig. 6.26 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the Full Buffer Traffic Type and the Windowing Factor of $\rho = 5.5$

is $\rho = 5.5$ (Fig. 6.26). ACLA and GPF-LM schemes achieve more than 45% of mean percentage of TTIs when all the active bearers are satisfied.

The percentages of TTIs for different reward types are analyzed in Figures 6.27, 6.28 and 6.29. ACLA, QVMAX and QV learning procedures show the highest amount of maximum rewards of about more than 7% for a windowing

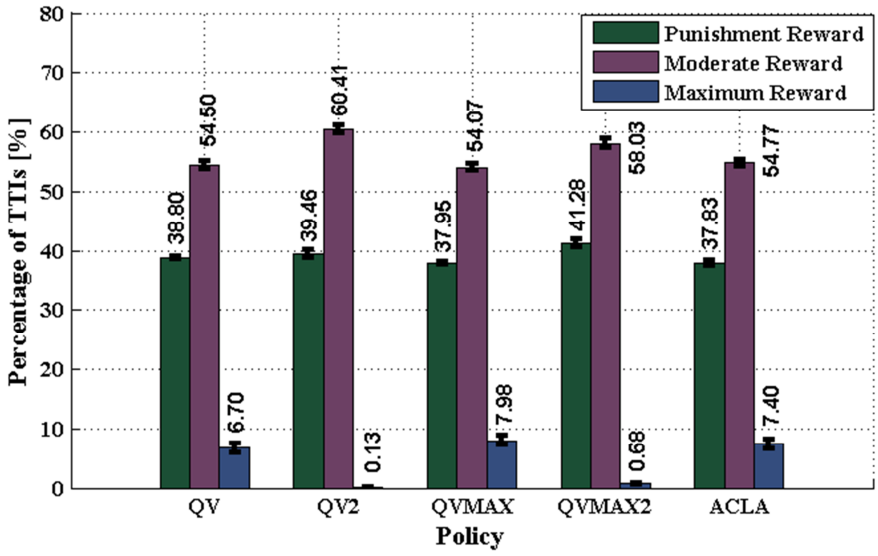


Fig. 6.27 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the Full Buffer Traffic Type and the Windowing Factor of $\rho = 2.5$

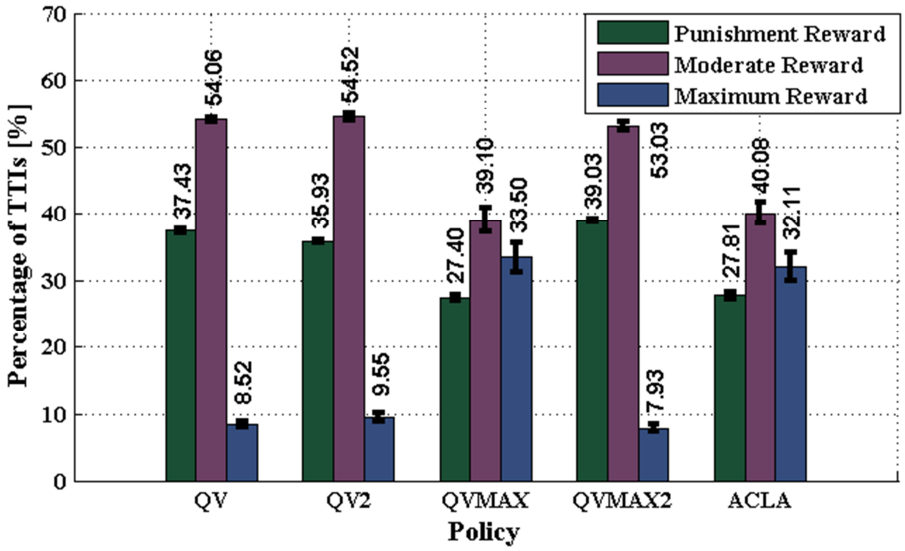


Fig. 6.28 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the Full Buffer Traffic Type and the Windowing Factor of $\rho = 4.0$

factor of $\rho = 2.5$. For higher factors (e.g., $\rho = 4.0$ and $\rho = 5.5$), the scheduling policy provided by the ACLA RL algorithm indicates the highest percentages of TTIs when the maximum rewards are considered in the exploitation period. It is important to notice that the filter length has a big impact in the DSR-SMOO problems since the amount of TTIs with punishment and moderate rewards decreases when the windowing factor increases (Figs. 6.27, 6.28 and 6.29).

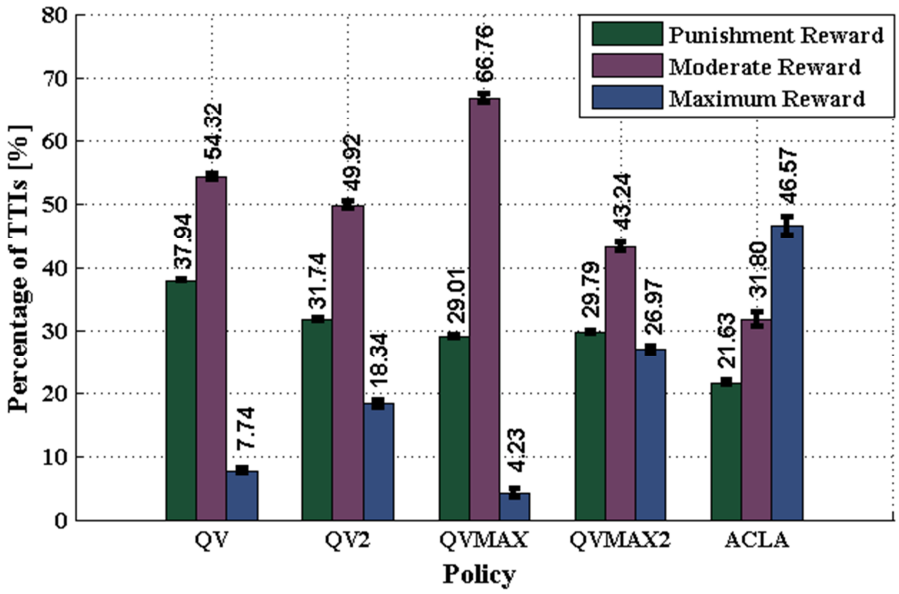


Fig. 6.29 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the Full Buffer Traffic Type and the Windowing Factor of $\rho = 5.5$

The same concept is studied in Fig. 6.30 where the percentages of TTIs for maximum, moderate and punishment rewards are analyzed for the ACLA actor-critic scheme. The percentage of TTIs with punishment rewards is reduced by about 20% for $\rho = 5.5$ when compared with the case when $\rho = 2.0$. This effect is more visible when the number of TTIs with the maximum rewards is counted by indicating a gain of 45% for $\rho = 5.5$ when compared with the case of $\rho = 2.0$. From Figure 6.30, it can be concluded that ACLA offers the best policy in terms of the number of TTI when the maximum testing reward is registered while QV2 learning indicates the best policy which is able to recover the GBR feasibility state by indicating the best percentage of TTIs of moderate rewards.

Figure 6.31 illustrates the comparison between ACLA actor-critic learning and the proposed scheduling rule GPF-LM from the viewpoints of the mean percentages of TTIs $\overline{p}_{TTI}^{G,100\%}$, $\overline{p}_{TTI}^{G,95\%}$, $\overline{p}_{TTI}^{G,90\%}$, $\overline{p}_{TTI}^{G,85\%}$ and $\overline{p}_{TTI}^{G,80\%}$ for a varying windowing factor in the case of the full buffer traffic type. Comparable results between the two technologies can be achieved for $\rho = \{3.0; 4.5; 5.5\}$ static values, which compute the GBR reward functions. When the windowing factor is $\rho = 2.0$, the GPF-LM scheduling metric outperforms with about 40% for $\overline{p}_{TTI}^{G,80\%}$

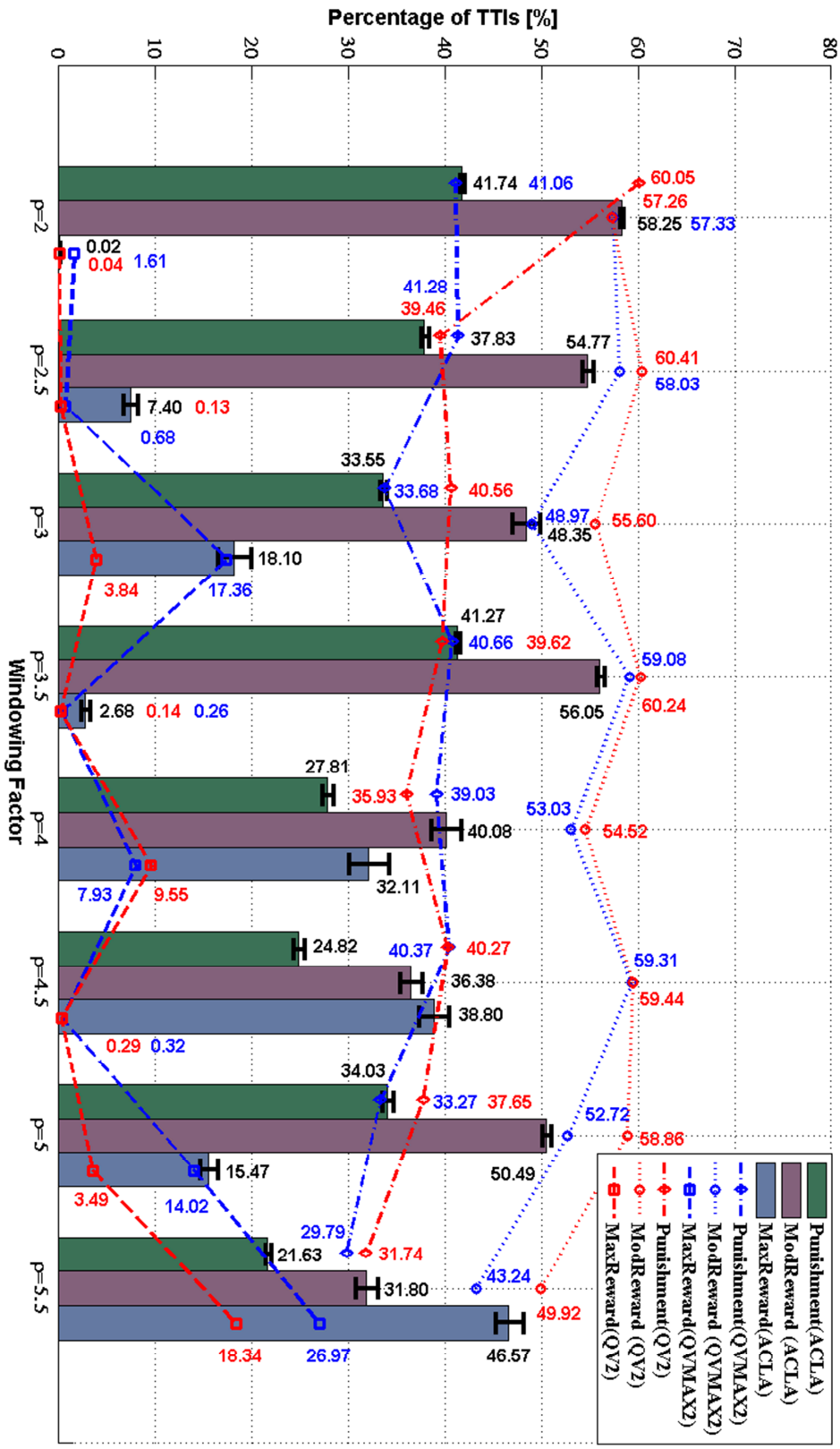


Fig. 6.30 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the Full Buffer Traffic Type and the Variable Windowing Factor

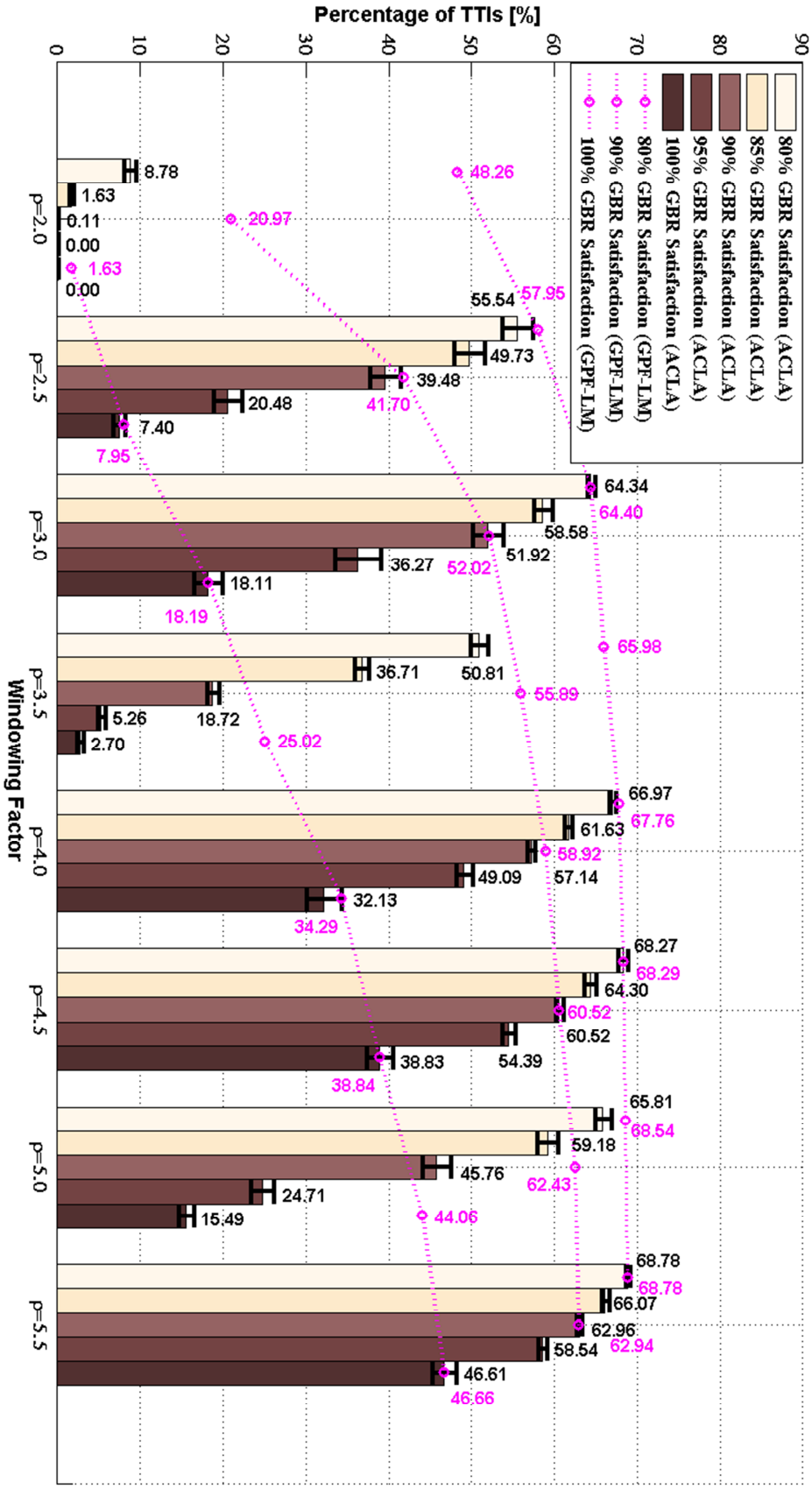


Fig. 6.31 Mean Percentages of TTIs for $\{100\%;95\%;90\%;85\%;80\%\}$ GBR Satisfaction with the Full Buffer Traffic Type and the Variable Windowing Factor

and with about 2% when the mean percentage of feasible TTIs $\overline{p}_{TTI}^{G,100\%}$ is considered, revealing the difficulty of the proposed policy to take optimal actions for very restrictive windowing factors. When the windowing factor increases to $\rho = 3.5$, the degradation of the mean percentage of feasible TTIs $\overline{p}_{TTI}^{G,100\%}$ for the ACLA RL algorithm, when compared with GPF-LM, is more than 22% which requires in fact more epochs of exploration for all considered RL algorithms. In general, it can be concluded that for the full buffer traffic type, the combination of multiple scheduling rules is not able to increase the percentage of TTIs with 100% satisfied bearers since the proposed GPF-LM scheduling rule performs the best for all possible scenarios.

A complete set of results for the DSR-SMOO MDP problems focusing on the GBR requirement for the infinite traffic type is listed in Appendix G, where Tables G.1 to G.8 analyze the performance of the obtained RL policies from the viewpoint of the mean percentage of TTIs where the GBR satisfaction domain is $\left[\overline{p}_{TTI}^{G,91\%}, \overline{p}_{TTI}^{G,100\%}\right]$. Tables G.25 to G.32 present the simulation results for the same scheduling policies when the mean percentages of TTIs with moderate and punishment rewards are analyzed for the considered domain of windowing factors. The scheduling policies obtained by using the ACLA actor-critic scheme are considered sustainable since the mean percentages of feasible TTIs are maximized, the number of punishments is minimized and the STD values for the performance indicators are minimized. The best performance is achieved for the optimum windowing factor of $\rho = 5.5$ when the mean percentage of TTIs with punishment rewards is minimized among the entire domain of windowing factors.

6.3.4.4 DSR-SMOO GBR with the CBR Arrival Rate

When the Constant Bit Rate (CBR) traffic type is considered, the impact of the optimization problem differs from the case of full buffer model since the data packets should be scheduled in the downlink sense in order to satisfy the stability condition for each MAC data queue $\overline{\overline{T}}_i[t] \le \overline{\lambda}_i[t]$. The similar scenario and the

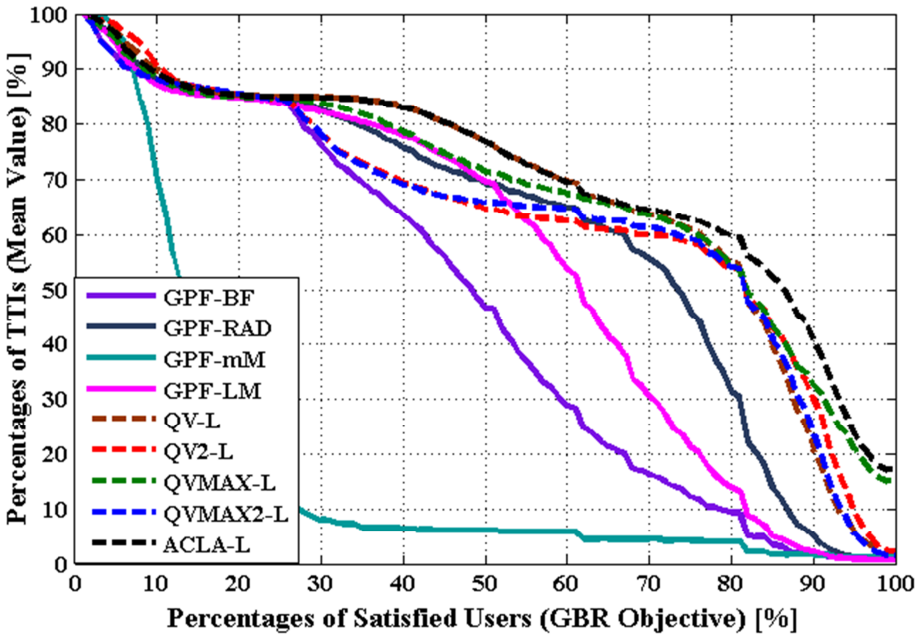


Fig. 6.32 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the CBR Traffic Type and the Windowing Factor of $\rho = 2.5$

performance indicators from the previous sub-section are reloaded in this sub-section with the amendment that the arrival rate is changing in accordance with the GBR constraint at each 1000 TTIs.

Figure 6.32 shows the performance of the proposed optimization model when the windowing factor is $\rho = 2.5$. The ACLA actor-critic scheme outperforms other approaches for almost the entire domain of the GBR satisfaction $\left[\overline{p}_{TTI}^{G,20\%}; \overline{p}_{TTI}^{G,100\%}\right]$. In particular, ACLA and QVMAX policies gain more than 15% percentage of TTIs when all users are satisfied, when compared against other scheduling rules or RL policies. When $\rho = 4.0$ (Fig. 6.33), the advantage of using the GBR DSR-SMOO optimization model becomes more visible due to the fact that the policy learned with ACLA scheme outperforms any of other scheduling rules of about 25% of TTIs when the level of GBR satisfaction is 100%. However, a performance degradation of QV and QVMAX schemes is registered when the GBR performance domain is $\left[\overline{p}_{TTI}^{G,20\%}, \overline{p}_{TTI}^{G,50\%}\right]$, since in this interval, only ACLA and QV2 policies are able to follow the GPF-LM scheduling rule. The same behavior of the GBR SMOO optimization problem

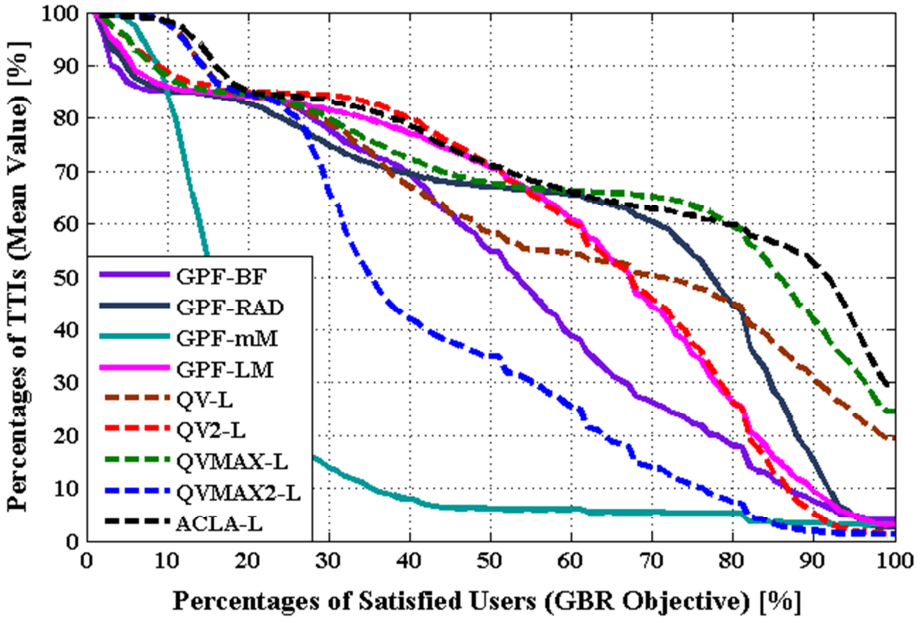


Fig. 6.33 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the CBR Traffic Type and the Windowing Factor of $\rho = 4.0$

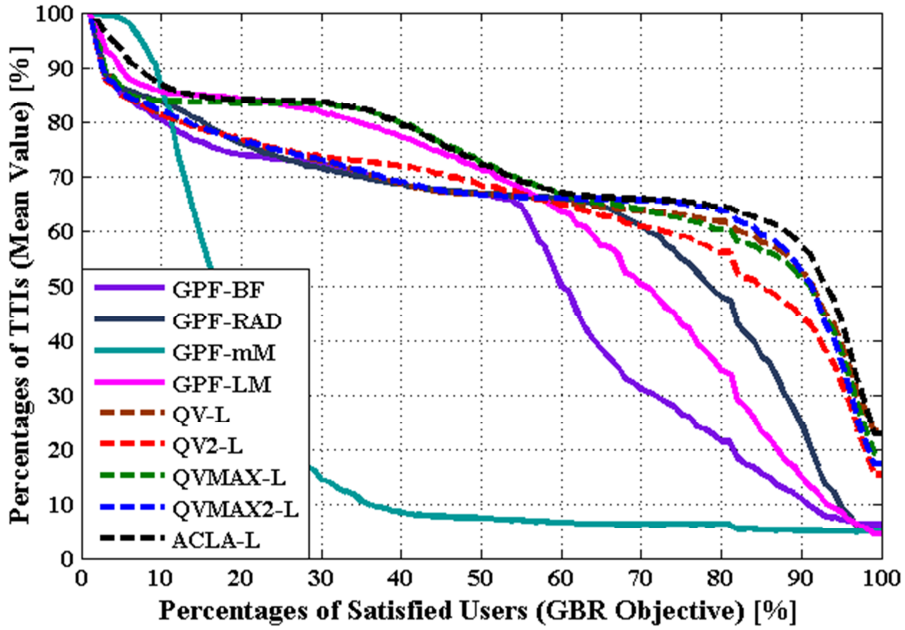


Fig. 6.34 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the CBR Traffic Type and the Windowing Factor of $\rho = 5.5$

is highlighted in Fig. 6.34 when $\rho = 5.5$. In this case, $\overline{p}_{TTI}^{G,100\%}$ is reduced with about 8% for the ACLA policy when compared with case of $\rho = 4.0$. All the considered RL approaches perform much better than other static scheduling rules

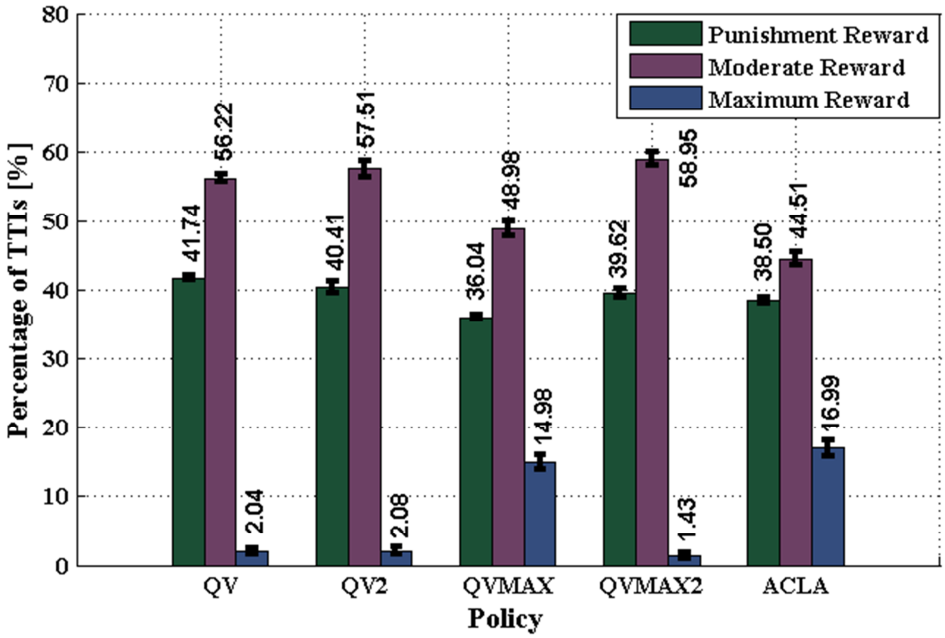


Fig. 6.35 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the CBR Traffic Type and the Windowing Factor of $\rho = 2.5$

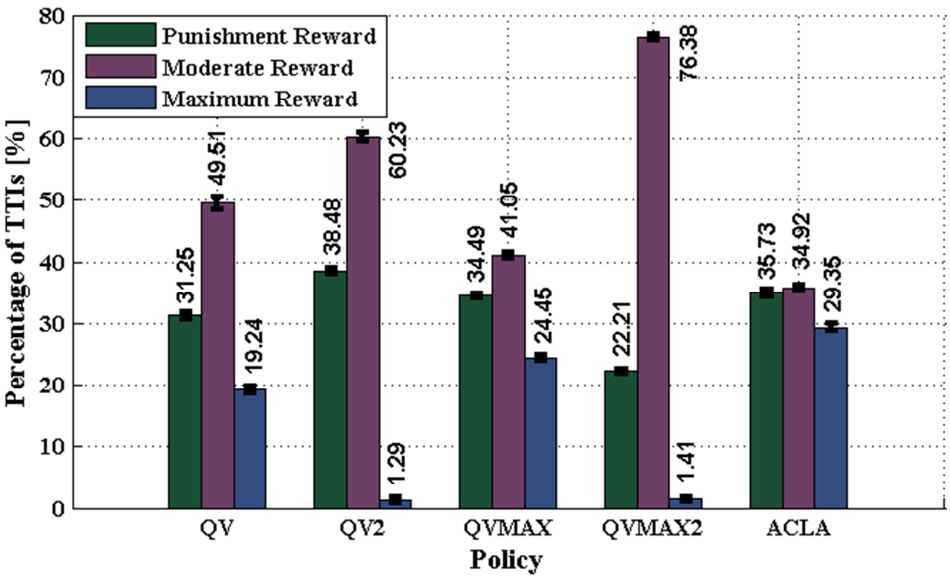


Fig. 6.36 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the CBR Traffic Type and the Windowing Factor of $\rho = 4.0$

when the $\left[\bar{p}_{TTI}^{G,70\%}, \bar{p}_{TTI}^{G,100\%}\right]$ GBR satisfaction domain is considered. Below this interval, only ACLA and QVMAX policies are able to indicate comparable GBR satisfaction levels. When the percentages of TTIs with maximum testing rewards are taken into account (Figs. 6.35, 6.36 and 6.37), the ACLA methodology shows

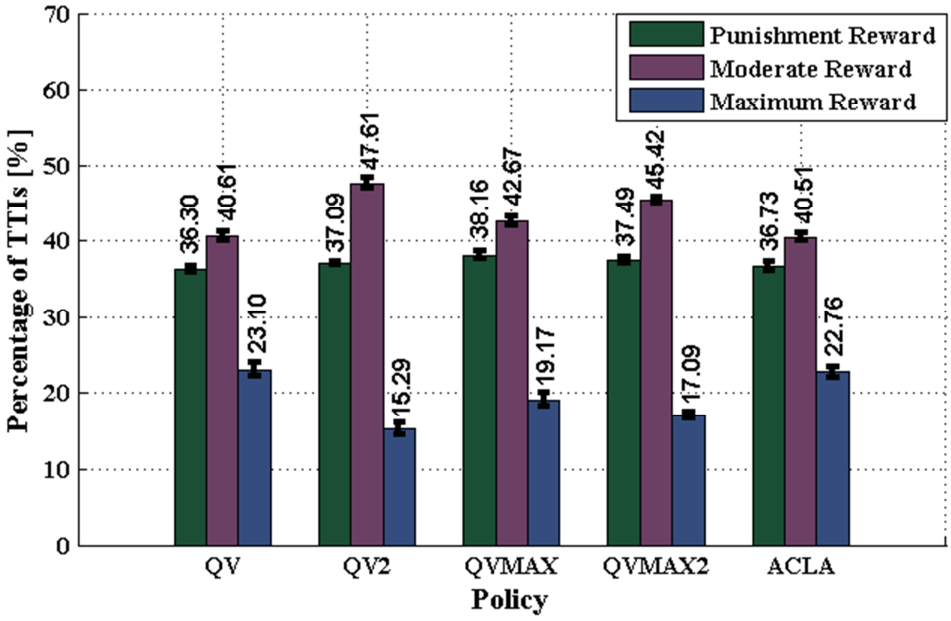


Fig. 6.37 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the CBR Traffic Type and the Windowing Factor of $\rho = 5.5$

the best performances excepting the case when $\rho = 5.5$. The maximum percentage value is achieved when the windowing factor of $\rho = 4.0$ (Fig. 6.36) is used to compute the reward function. By using a windowing factor of $\rho = 5.5$, the QV scheduling policy indicates a slight improvement from the percentage of maximum rewards when compared with the ACLA actor-critic method.

The percentages of reward type evolution for different windowing factors are presented in Fig. 6.38. As mentioned above, the best performance from the viewpoints of the mean percentages of TTIs $\left\{ \overline{p}_{TTI}^{G,PSH}, \overline{p}_{TTI}^{G,mRW}, \overline{p}_{TTI}^{G,MRW} \right\}$ is obtained when the windowing factor is $\rho = 4.0$ which is considered to be the optimum value for the CBR traffic type when the GBR DSR-SMOO MDP problem is performed. QVMAX policy assures the highest amount of moderate rewards by decreasing the amount of punishments, but unfortunately, without significant results from the viewpoint of the maximum rewards.

When the mean percentages of TTIs for $\left\{ \overline{p}_{TTI}^{G,100\%}; \overline{p}_{TTI}^{G,95\%}; \overline{p}_{TTI}^{G,90\%}; \overline{p}_{TTI}^{G,85\%}; \overline{p}_{TTI}^{G,80\%} \right\}$ GBR satisfaction levels are considered (Fig. 6.39), the GPF-LM scheduling rule presents degraded performances when compared with the ACLA policy for the

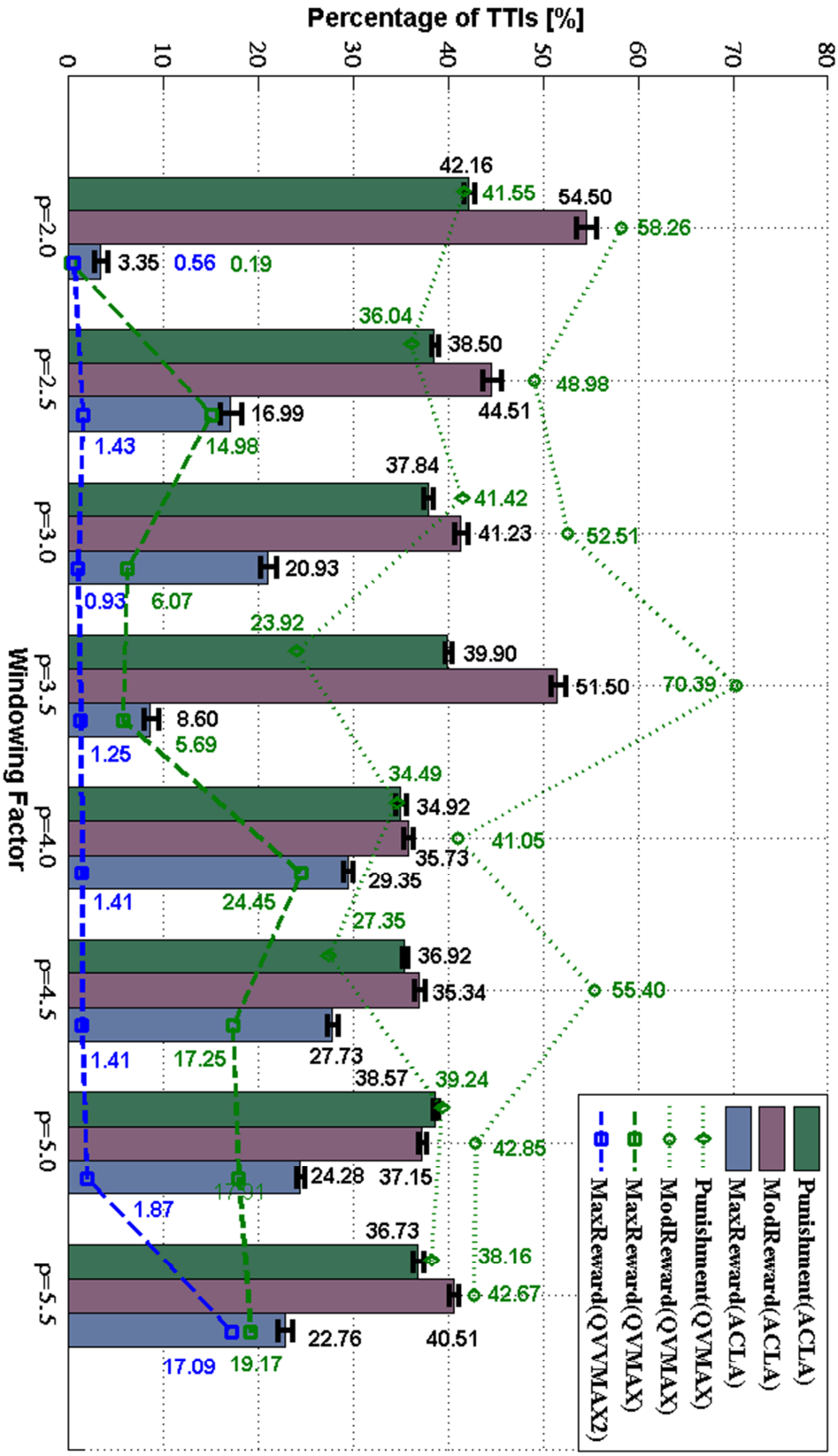


Fig. 6.38 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the CBR Traffic Type and Variable Windowing Factors

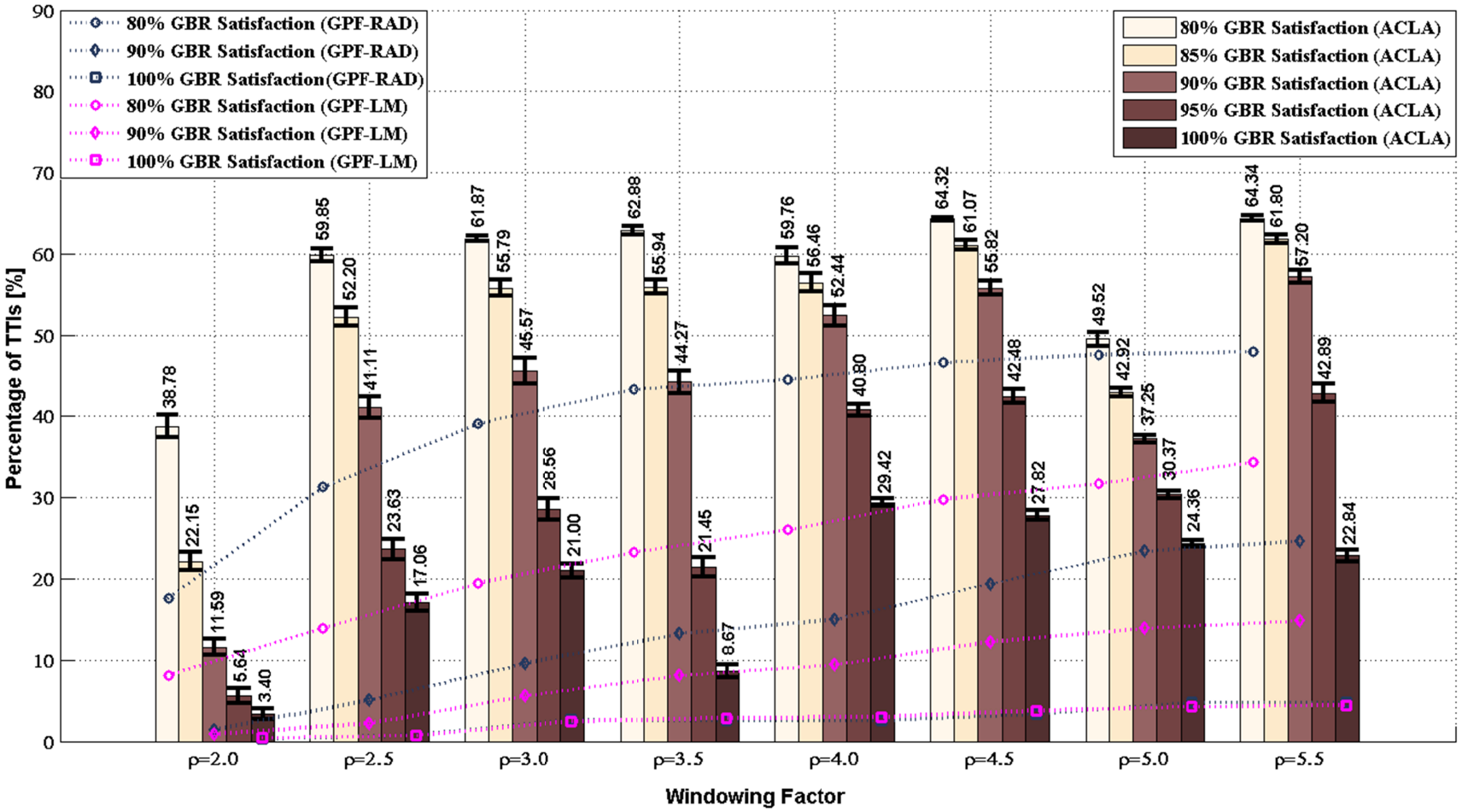


Fig. 6.39 Mean Percentages of TTIs for $\{100\%;95\%;90\%;85\%;80\%\}$ GBR Satisfaction with the CBR Traffic Type and Variable Windowing Factors

considered domain of windowing factors. When $\overline{p}_{TTI}^{G,90\%}$ and $\overline{p}_{TTI}^{G,80\%}$ are measured, the GPF-RAD rule outperforms the proposed scheduling scheme GPF-LM. The optimum filter length for ACLA considers a windowing factor of $\rho = 4.0$ when all active bearers are satisfied. For a lower percentage of GBR satisfaction, a higher windowing factor can be used as reference values (e.g. $\rho = 5.5$, Fig. 6.39).

In Appendix G, Tables G.9 to G.16 and Tables G.33 to G.40 extend the simulation results provided in this sub-section for the entire set of scheduling policies being obtained with other RL approaches. Different RL approaches for different windowing factor settings are the best choices from the viewpoint of the mean percentage of feasible TTIs. But in all the cases, the STD values for these performance indicators are minimized which reflect in fact, the sustainability of the proposed scheduling policies.

6.3.4.5 DSR-SMOO GBR with the VBR Arrival Rate

For the Variable Bit Rate (VBR) traffic type, the packet sizes and arrival rates are modeled by using the Pareto and geometric distributions, respectively, with exactly the same random variables at each TTI on the exploitation stage for the considered scheduling candidates oriented on the GBR objective. Since the arrival rate is not associated with the GBR constraints anymore, the optimum windowing factor should be larger in order to assure the scheduler stability.

The scheduling policies obtained by using QV2, ACLA and QV RL algorithms perform better than other approaches when the GBR satisfaction domain of $\left[\overline{p}_{TTI}^{G,90\%}, \overline{p}_{TTI}^{G,100\%}\right]$ is considered for a reference windowing factor of $\rho = 2.5$ (Fig. 6.40). QV and QVMAX select the best scheduling rules which are able to increase the number of satisfied bearers when the GBR satisfaction domain $\left[\overline{p}_{TTI}^{G,10\%}, \overline{p}_{TTI}^{G,80\%}\right]$ is analyzed. It is important to remind that the main purpose of the optimization problem $\left(P_G\right)$ is to increase the mean percentage of TTIs when all the active bearers are satisfied from the GBR objective point of view $\left(\overline{p}_{TTI}^{G,100\%}\right)$.

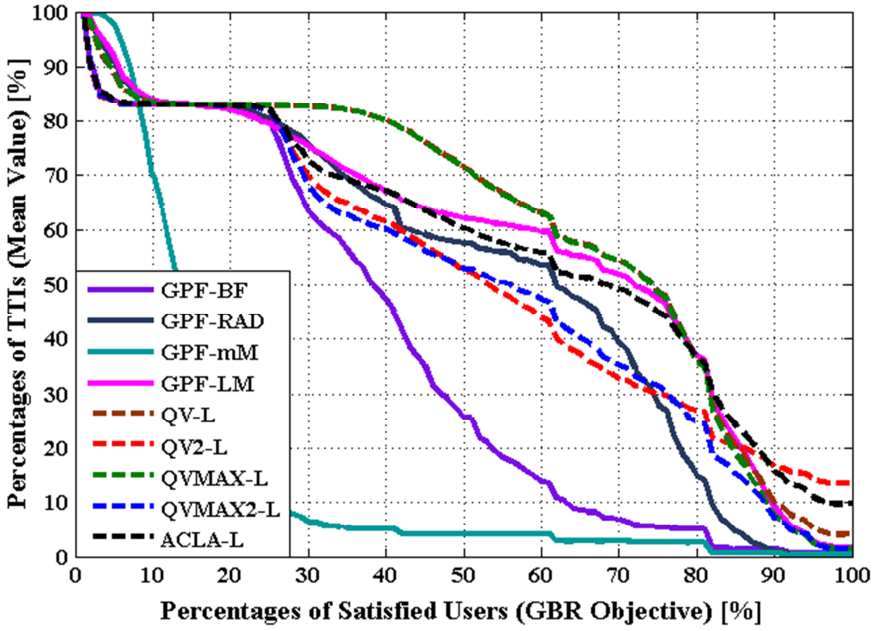


Fig. 6.40 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the VBR Traffic Type and the Windowing Factor of $\rho = 2.5$

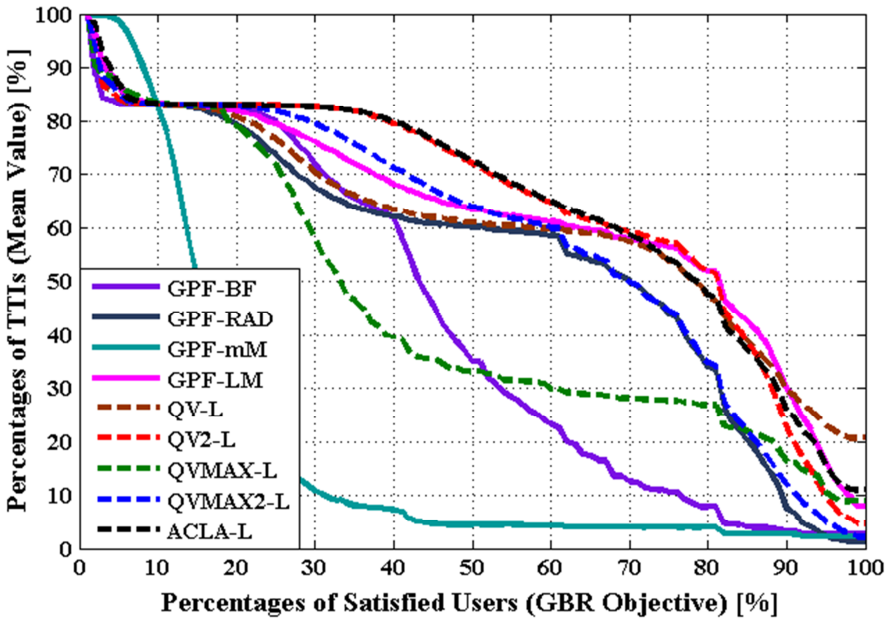


Fig. 6.41 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the VBR Traffic Type and the Windowing Factor of $\rho = 4.0$

When the reference factor is $\rho = 4.0$ (Fig. 6.41), the QV learning shows significant improvements of about 10% when compared with other candidates. For a GBR satisfaction in the domain of $\left[\overline{p}_{TTI}^{G,80\%}; \overline{p}_{TTI}^{G,90\%} \right]$, the GPF-LM discipline

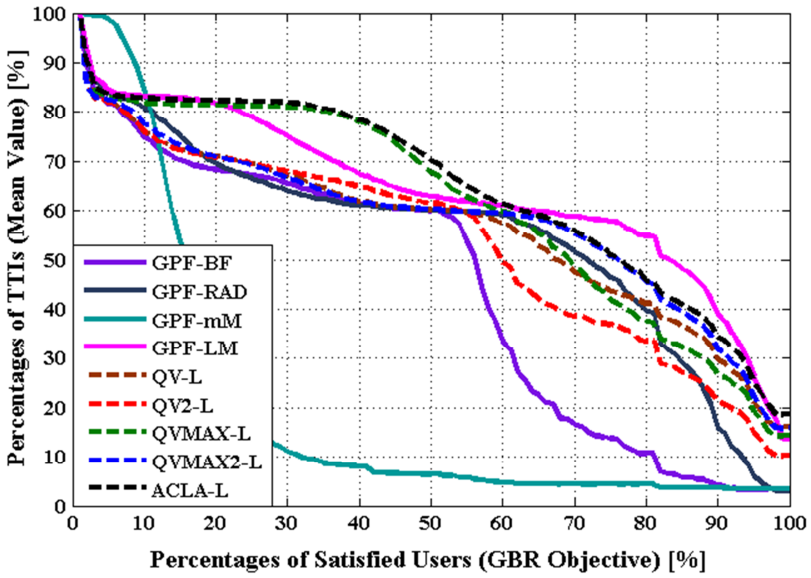


Fig. 6.42 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the VBR Traffic Type and the Windowing Factor of $\rho = 5.5$

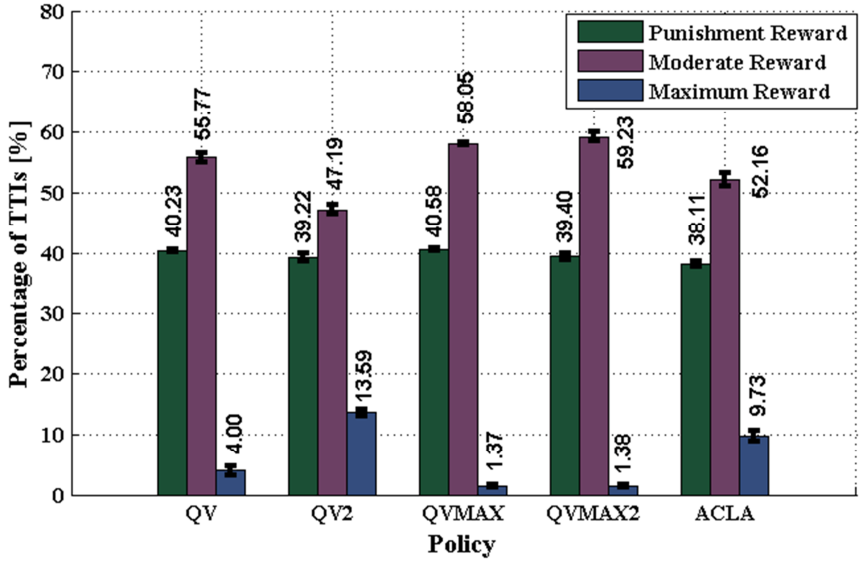


Fig. 6.43 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the VBR Traffic Type and the Windowing Factor of $\rho = 2.5$

offers the best performance when matched against other static scheduling rules or scheduling policies. The same GPF-LM rule performs the best in the domain of $\left[\bar{p}_{TTI}^{G,65\%}; \bar{p}_{TTI}^{G,95\%}\right]$ when the windowing factor is $\rho = 5.5$ (Fig. 6.42). Excepting this interval, ACLA policy remains the best option. From the viewpoint of the mean percentages of TTIs with maximum rewards $\bar{p}_{TTI}^{G,MRW}$, the entire amount of

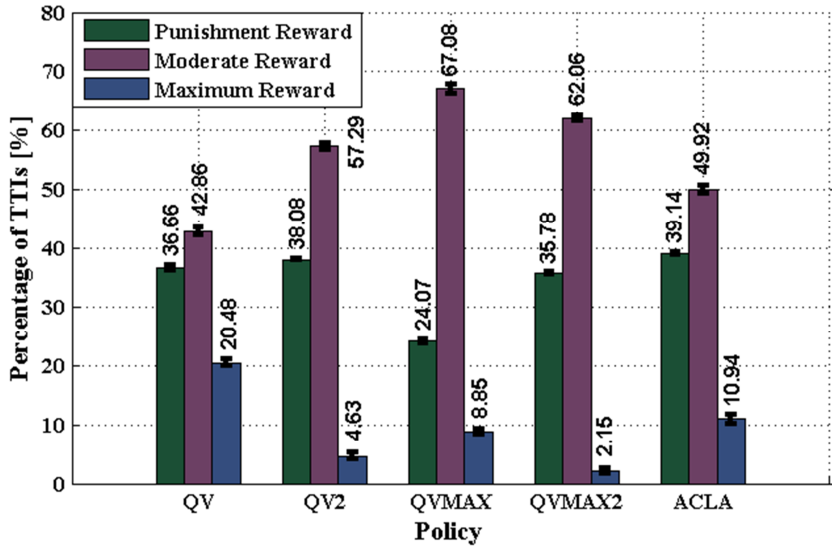


Fig. 6.44 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the VBR Traffic Type and the Windowing Factor of $\rho = 4.0$

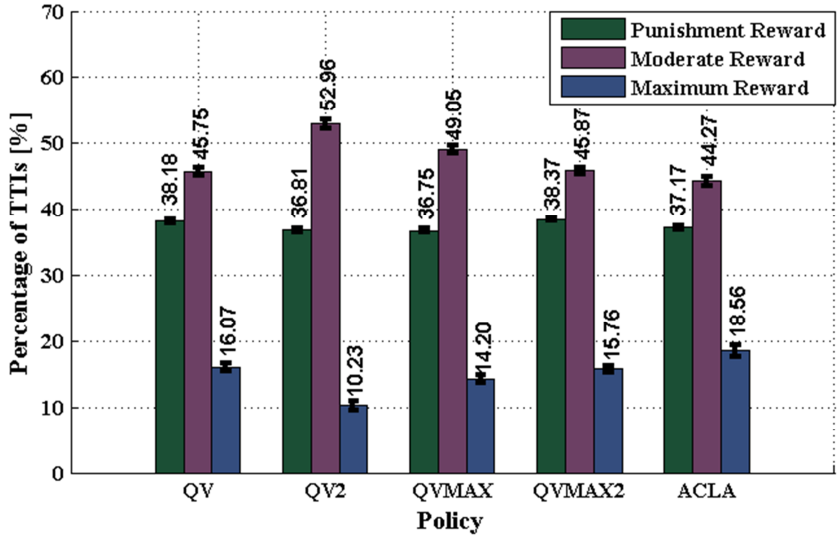


Fig. 6.45 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the VBR Traffic Type and the Windowing Factor of $\rho = 5.5$

maximum rewards for the windowing factors of $\rho \in \{2.5; 4.0; 5.5\}$ is reduced when compared with the CBR and full buffer traffic types (Figs. 6.43, 6.44, 6.45). The feasibility state is reached more difficult for the VBR traffic than for any of other analyzed traffic models. In this sense, a larger windowing factor is required in order to model the DSR-SMOO MDP problem (P_G) as an episodic task. For instance, ACLA policy achieves a level of $\overline{p}_{TTI}^{G,MRW} = 18.56\%$ maximum rewards

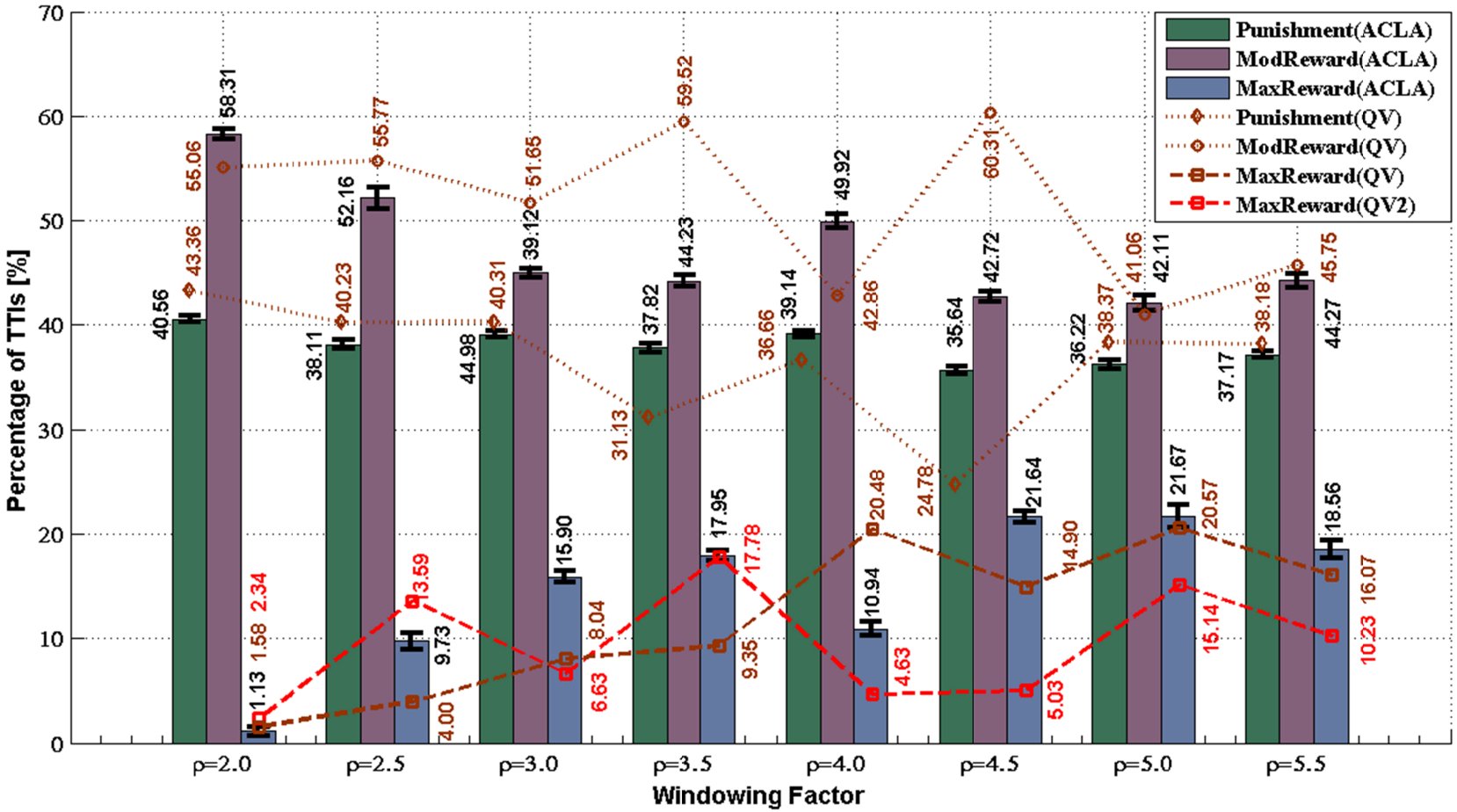


Fig. 6.46 Mean Percentages of TTIs for Punishment, Moderate and Maximum Rewards with the VBR Traffic Type and Variable Windowing Factors

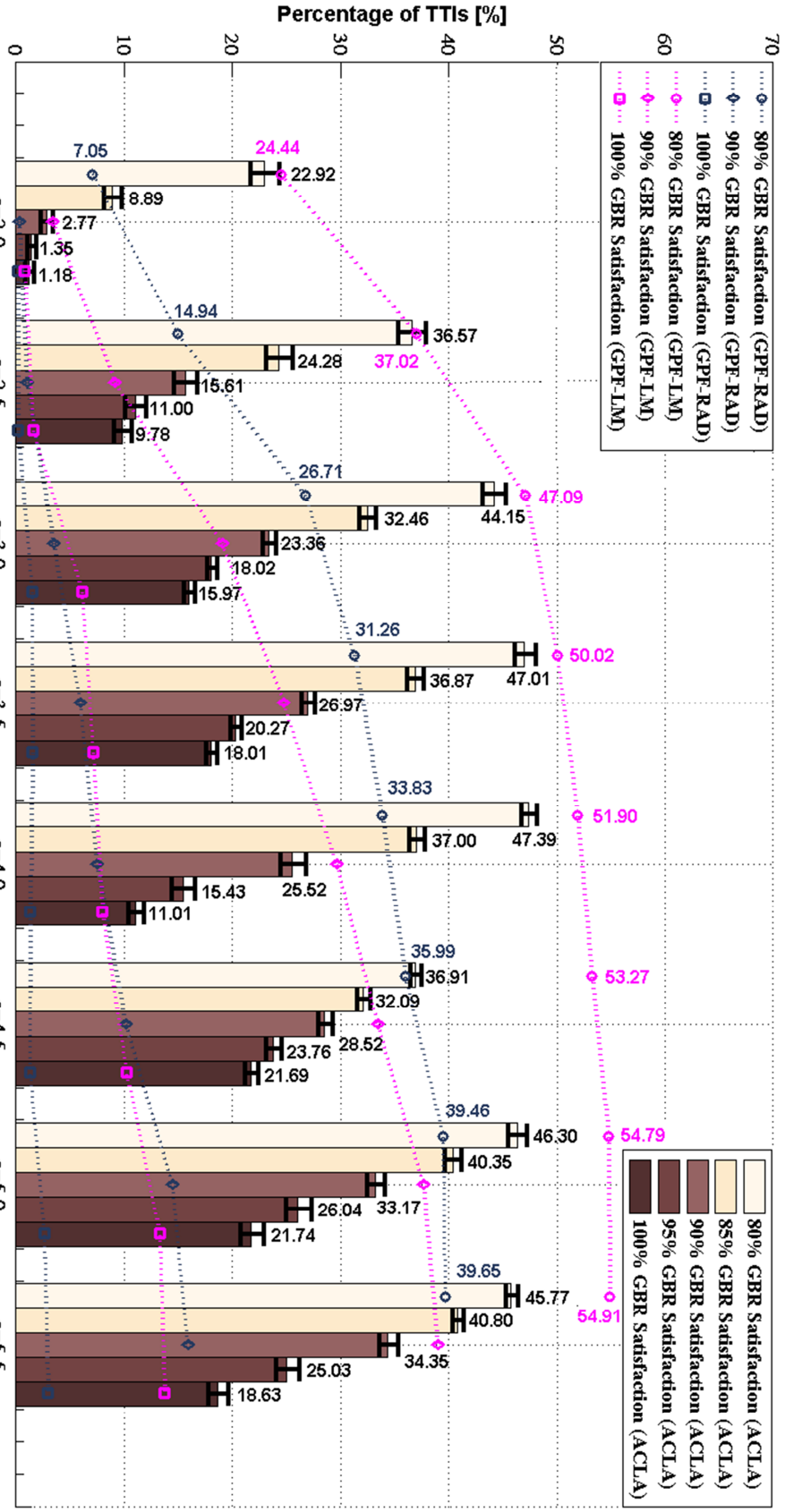


Fig. 6.47 Mean Percentages of TTIs for $\{100\%;95\%;90\%;85\%;80\%\}$ GBR Satisfaction with the VBR Traffic Type and Variable Windowing Factors

when the windowing factor is $\rho = 5.5$ (Fig. 6.45). An exception is represented by the QV policy which achieves a level of $\overline{p}_{TTI}^{G,MRW} = 20.48\%$ (Fig. 6.44) when the windowing factor is $\rho = 4.0$, performance which is comparable with the QV policy for the CBR traffic type for the same windowing factor.

In Fig. 6.46, the evolution of $\left(\overline{p}_{TTI}^{G,PSH}, \overline{p}_{TTI}^{G,mRW}, \overline{p}_{TTI}^{G,MRW}\right)$ is depicted for the ACLA actor-critic method in order to localize the optimum windowing factor. As seen from Fig. 6.46, the amount of moderate rewards decreases when the windowing factor increases by indicating a minimum percentage of TTIs when the windowing factor is $\rho = 5.0$. From the perspective of percentage $\overline{p}_{TTI}^{G,MRW}$, the same factor for the filter length is considered as an optimum value. The optimum windowing factor for the VBR traffic type belongs to $\rho \in [4.5; 5.0]$ even if the amount of punishments when $\rho = 5.0$ is slightly higher than the case of $\rho = 4.5$. More details can be found in Tables G.41-G.48 from Appendix G.

From the mean percentage $\overline{p}_{TTI}^{G,100\%}$ point of view, ACLA performs the best for the considered windowing factors as shown by Fig. 6.47 when compared with other static scheduling rules. The main candidate, GPF-LM increases the mean percentages of TTIs for the GBR satisfaction levels of $\overline{p}_{TTI}^{G,90\%}$ and $\overline{p}_{TTI}^{G,80\%}$ when compared with ACLA policy and the GPF-RAD scheduling rule. Despite these aspects, ACLA remains the best choice from the $\overline{p}_{TTI}^{G,100\%}$ percentage point of view. A complete set of results for the discussed policies and for other considered scheduling policies is presented in Appendix G in Tables G.17 to G.24. The best scheduling policies being oriented on the GBR objective for the VBR traffic type are sustainable due to the facts that the number of feasible TTIs is maximized and the number of punishments is minimized especially for higher windowing factors.

6.4 Summary

Two types of sequential optimization problems have been addressed in this chapter in terms of the NGMN fairness and GBR objectives. It has been shown

that the type of filter used in averaging the user throughputs plays a crucial role in achieving the NGMN fairness or the GBR targets. The AUT-EMF observations have been used in order to show the utility of using the aggregate CQI information in the controller state space computation. By using different CQI aggregation schemes, CACLA2 and CACLA1 policies gain more than 11% feasible TTIs when compared with the existing approaches and with other policies which do not consider the CQI aggregation principle. When the AUT-MMF observations are used, the windowing factor and the maximum number of schedulable users at each TTI have a great impact in the filter window length computation. By exploiting the CACLA1 or CACLA2 scheduling policies, it has been registered a gain in the percentages of TTIs when the scheduler is feasible from the NGMN requirement point of view, when compared with the existing approaches, of about 15% to 35% for different time windowing settings. Also, the percentage of TTIs when the scheduler is considered unfair is minimized with about 16% to 25% for the same range of widowing factors.

When the sequential optimization focusing on the GBR performance is considered, alongside the windowing factor parameterization, the type of simulated traffic has a great influence on the performance of the exploited policies. In this sense, the proposed scheduling rule GPF-LM is able to outperform other classical metrics from the percentages of TTIs when the active bearers are satisfied in proportion of 100% for the full buffer traffic model. When the DSR-SMOO problem is performed for the same type of traffic, only ACLA actor-critic scheme can follow the trajectory imposed by GPF-LM by indicating a gain of 5% to 40% if the 100% GBR satisfaction level is considered when compared against the existing methods excepting GPF-LM. With the CBR traffic type, the combination of different scheduling rules improves the GBR satisfaction of active bearers by about 15% to 20% when the ACLA policy is exploited. When the VBR traffic type is scheduled, the GPF-LM increases the percentage of TTIs in the domain of $\left[\overline{p}_{TTI}^{G,65\%}; \overline{p}_{TTI}^{G,95\%}\right]$ for large windowing factors. Otherwise, ACLA, QV and QV2 learning procedures show the best performance, especially when the mean percentage of TTIs in the interval of $\left[\overline{p}_{TTI}^{G,95\%}, \overline{p}_{TTI}^{G,100\%}\right]$ is analyzed.

Chapter 7

Sustainable Scheduling Policies for Concurrent Multi-Objective Optimization

7.1 Chapter Outline

The concurrent optimization aims to merge the QoS reward functions in order to learn the sustainable policies which can select at each TTI optimum scheduling rules focusing on different scheduling objectives. This chapter analyzes two types of concurrent optimizations. The first DSR-CMOO MDP problem refers to the HoL delay and PDR multi-objective evaluation which strongly depends on the windowing factor used for the PDR computation. At the same time, the particularities of each scheduling rule are studied in order to highlight the importance of each discipline in a given HoL delay and PDR tradeoff performance domain. The second DSR-CMOO MDP problem includes four objectives of NGMN user fairness, GBR, HoL packet delay and PDR requirements. An enhanced version of CACLA2 is used in this sense in order to find the optimum windowing factor at different time intervals for the NGMN fairness, GBR and PDR objectives. Based on the novel RL approach, the obtained sets of scheduling policies are able to perform much better than other standard scheduling rules by maximizing the mean percentage of feasible TTIs and by minimizing at the same time, the amount of punishment rewards in the exploitation stage when the considered tradeoff is analyzed.

7.2 DSR-CMOO MDP Focusing on HoL Packet Delay and PDR Objectives

The DSR-CMOO MDP problems focusing on HoL packet delay and PDR objectives select at each TTI the scheduling rule which can provide the highest scheduler reward in terms of the merged delay and PDR rewards. Similar to the GBR reward function, the HoL delay and PDR rewards are computed based on the delay and drop rate requirements. The data packets which exceed the HoL delay requirements ($d_i^{\overline{HoL}}[t]$) are automatically dropped in order to permit other packets that approach to the deadline to be scheduled. In this sense, the HoL packet delay objective is satisfied at each TTI when the packet drop module is used, by degrading at the same time the PDR performance. The packet drop rate can be decreased if the packet delay budget is managed properly by imposing a lower delay constraint ($d_{i,L}^{\overline{HoL}}[t] << d_i^{\overline{HoL}}[t]$) to be satisfied at each TTI. Another factor which has a great impact on the PDR objective refers to the time window length which is used to calculate the packet drop rate. Therefore, the merged Delay and PDR (DP) reward function should minimize the mean HoL packet delay and decrease the PDR rate for a given time window.

The mean HoL delay $\left(\overline{d^{HoL}}[t] = \sum_{i=1}^{|\mathcal{U}_t|} d_i^{HoL}[t] / |\mathcal{U}_t|\right)$ performance can be minimized by using two characteristics which affect the scheduler's feasibility:

1. When the standard deviation of HoL delays is very close to its mean value and $\overline{d^{HoL}}[t] < d_{i,L}^{\overline{HoL}}[t]$, the state optimality in terms of the HoL delay ($d_i^{HoL}[t] < d_{i,L}^{\overline{HoL}}[t],\ \forall i \in \mathcal{U}_t$) can be reached much faster.
2. When the standard deviation of HoL delays is relatively higher and $\overline{d^{HoL}}[t] < d_{i,L}^{\overline{HoL}}[t]$, the optimality of the HoL packet delay for all active bearers is not guaranteed and a given percentage of active bearers can be in outage from the $d_{i,L}^{\overline{HoL}}$ lower constraint point of view.

The idea of the learned concurrent scheduling policies is to avoid applying scheduling rules which minimize the mean delay $\overline{d^{HoL}}[t]$ when a high STD value is involved. Therefore, the aim of this section is to propose a set of scheduling policies which is able to minimize the packet drop rate and to minimize at the same time, the mean and STD values for the HoL packet delays at each TTI.

7.2.1 The Optimization Problem

The considered scheduling optimization problem includes four scheduling rules focusing on the HoL packet delay being introduced in Chapter 3, namely, GPF-EDF, GPF-LOG, GPF-EXP1 and GPF-EXP2. The purpose of the DSR-CMOO problem is to obtain a set of scheduling policies which applies one of the proposed scheduling rules at each TTI in order to respect the $\overline{d_{i,L}^{HoL}}[t]$ and $\overline{R_i^{PL}}[t]$ constraints TTI-by-TTI. The proposed DSR-CMOO problem focusing on packet delay and packet drop rate is highlighted by Eq. 7.1.a, where the fairness parameters are fixed to $(\alpha=1,\beta=1)$ and $\{c_{4,1}[t],c_{4,2}[t],c_{4,3}[t],c_{4,4}[t]\}$ represents the mapped controller actions responsible in selecting a given scheduling rule, and the MU assignation vectors are $\{u_{1,i}^4[t],u_{2,i}^4[t],u_{3,i}^4[t],u_{4,i}^4[t]\}$ which permit to select only one MU function, the same for each user $i\in\mathcal{U}_t$, at each TTI t. Based

$$
\begin{aligned}
(P_D): \max_{\pi_{RB}[t]} \Bigg\{ & c_{4,1}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{1,i}^4[t]\cdot b_{i,j}[t]\cdot\frac{1}{\overline{d_{i,L}^{HoL}}[t]-d_i^{HoL}[t]}\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right] 1:(GPF-EDF) \\
& +c_{4,2}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{2,i}^4[t]\cdot b_{i,j}[t]\cdot exp\left(\frac{\omega_{2,i}^4\cdot d_i^{HoL}[t]}{1+\sqrt{\widehat{d_{ik}^{HoL}}[t]}}\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right] 2:(GPF-EXP1) \\
& +c_{4,3}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{3,i}^4[t]\cdot b_{i,j}[t]\cdot exp\left(\frac{\omega_{3,i}^4\cdot d_i^{HoL}[t]-\widehat{d^{HoL}}[t]}{1+\sqrt{\widehat{d^{HoL}}[t]}}\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right] 3:\begin{pmatrix}GPF\\EXP2\end{pmatrix} \\
& +c_{4,4}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{4,i}^4[t]\cdot b_{i,j}[t]\cdot log\left(\omega_{4,1}^4+\omega_{4,2}^4\cdot\frac{d_i^{HoL}[t]}{\overline{d_{i,L}^{HoL}}[t]}\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right] 4:\begin{pmatrix}GPF\\LOG\end{pmatrix}
\end{aligned}
\tag{7.1.a}
$$

$$
\begin{aligned}
& c_{4,1}[t]+c_{4,2}[t]+c_{4,3}[t]+c_{4,4}[t]=1 \\
& \sum_{w_4=1}^{4} u_{w_4,i}^{4}[t]=1,\ i=1,...,|\mathcal{U}_t| \\
& \sum_{i=1}^{|\mathcal{U}_t|} u_{w_4,i}^{4}[t]=|\mathcal{U}_t|,\ w_4^{*}\in\mathcal{PU}_4 \\
(C_D)\ s.t.: & \sum_{i=1}^{|\mathcal{U}_t|} u_{w_4^{\otimes},i}^{4}[t]=0,\ w_4^{\otimes}=1,...,|\mathcal{PU}_4|,\forall w_4^{\otimes}\neq w_4^{*} \\
& \sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t]=1,\ j=1,...,|\mathcal{B}| \\
& b_{i,j}[t]\in\{0,1\},\ \forall i\in\mathcal{U}_t,\forall j\in\mathcal{B} \\
& \{u_{1,i}^{4}[t],u_{2,i}^{4}[t],u_{3,i}^{4}[t],u_{4,i}^{4}[t]\}\in\{0,1\},\forall i\in\mathcal{U}_t \\
& \{c_{4,1}[t],c_{4,2}[t],c_{4,3}[t],c_{4,4}[t]\}\in\{0,1\}
\end{aligned}
\tag{7.1.b}
$$

on the DSR-CMOO problems focusing on HoL packet delay and PLR objectives, the set of objective constraints (O_D) must be satisfied at each TTI *t* for each user $i\in\mathcal{U}_t$ as shown in Equation 7.1.c:

$$
\begin{aligned}
(O_D): \quad & d_i^{HoL}[t]\leq \overline{d_{i,L}^{HoL}}[t],\forall i\in\mathcal{U}_t \\
& R_i^{PL}[t]\leq \overline{R_i^{PL}}[t],\forall i\in\mathcal{U}_t
\end{aligned}
\tag{7.1.c}
$$

If the multi-objective constraints of Equation 7.1.c are satisfied for each active user $i\in\mathcal{U}_t$ at TTI *t*, then the feasible state is reached from the viewpoint of combined HoL delay and PDR objectives. The idea is to find this state as often as possible and to maximize the percentage of TTIs when all active users are satisfied from the aforementioned combined criterion.

Based on Equations 7.1.a, 7.2.b and 7.2.c, the role of the LTE scheduler controller is to approximate at each TTI *t* the optimum discrete action $\mathcal{A}_t^{a,DP}=\{1:(c_{4,1}[t]=1),2:(c_{4,2}[t]=1),3:(c_{4,3}[t]=1),4:(c_{4,4}[t]=1)\}$ based on the aggregate controller state information which is able to maximize the optimization problem (P_D) by respecting the set of convex constraints (C_D) and to satisfy the considered objectives (O_D) for all active bearers at each TTI. For example, if the controller selects the action $\mathcal{A}_t^{1,DP}=1$, then the selected scheduling rule is GPF-

EDF, if $\mathcal{A}_t^{2,DP} = 2$, the scheduling rule which has to be applied at the current TTI is GPF-EXP1 or if $\mathcal{A}_t^{4,DP} = 4$, then the applied rule is GPF-LOG.

As mentioned earlier, the main role is to increase the mean percentage of TTIs when all the considered bearers are 100% satisfied from the viewpoint of DP objectives and to minimize at the same time, the number of punishment and moderate testing rewards in the exploitation stage. As seen in Chapter 6, the idea is to exploit the learned scheduling policies by using different channel conditions. The mean percentage of DP feasible TTIs should be maximized and the STD values for the entire DP evaluation domain must be minimized in order to prove the sustainability of the obtained policies.

As seen from Eq. 7.1, the lower delay constraint $d_{i,L}^{\overline{HoL}}[t]$ is preferred in the detriment of the original constraint imposed by the 3GPP $d_i^{\overline{HoL}}[t]$ in the marginal utility function computation. Let us define the lower HoL packet delay requirement as indicated in Eq. 7.2:

$$d_{i,L}^{\overline{HoL}}[t] = \vartheta_D \cdot d_i^{\overline{HoL}}[t] \tag{7.2}$$

where, $\vartheta_D \in \mathbb{R}_{[0,1]}$ is the fraction of the LTE HoL delay requirement which is used in the proposed DSR-CMOO problem. Similar to the median filter length, the delay requirement parameter influences the number of episodic tasks during the exploration period. When ϑ_D is very low for a delay constraint of $d_i^{\overline{HoL}}[t] = 50ms$ and when a large number of users is considered, the DSR-CMOO MDP problems focusing on DP multi-objective are not episodic. Then, a special care should be given when the delay fraction parameter is set since it affects the mean percentage of TTIs when the scheduler is feasible from the viewpoint of DP objectives.

Another factor which impacts the number of episodic tasks for the DP CMOO optimization problem is the time window $\left(T_w^{PDR}\right)$ which is used for the packet drop rate computation at each TTI. From the operator point of view, the parameter T_w^{PDR} should be as large as possible in order to satisfy the PDR

objective for a large number of TTIs. From the LTE controller perspective, higher T_w^{PDR} parameters imply weak instantaneous PDR rewards which attract the sub-optimal action selection. In general, it is preferable to set $T_w^{PDR} = T_w^M$ being equal with the filter length which is used in the AUT-MMF computations, since the average user throughput and the packet drop rate should be considered in the MUTI computation by using the same time window. The PDR objective evaluation can be achieved by using a larger time window different from the one which is involved in the scheduling rule computation. Then, the rate of lost packets can be calculated based on Eq. 7.3:

$$R_i^{PL}[t] = \left(\sum_{w=t}^{T_w^{PDR}} \mathcal{N}_i^{LP}[w-t] \right) \Big/ \left(\sum_{w=t}^{T_w^{PDR}} \mathcal{N}_i^{TP}[w-t] \right) \tag{7.3}$$

where $\mathcal{N}_i^{LP}[t]$ denotes the number of lost packets at TTI t by user $i \in \mathcal{U}_t$, and $\mathcal{N}_i^{TP}[t]$ represents the total number of transmitted packets (with ACK acknowledge) for user $i \in \mathcal{U}_t$ at TTI t. The packets are declared lost if a NACK message is received or if the packets are declared dropped when the real HoL delay requirement is exceeded. For this study, *only the number of dropped packets is considered* in the number of lost packets $\mathcal{N}_i^{LP}[t]$ online computation.

7.2.2 Controller State Space

In order to describe the controller state elements, the feasibility and the unfeasibility in terms of the DP multi-objective criterion should be defined. Let us define $\mathcal{S}_t^{C,D} \in \mathcal{UFD}$ the unfeasible regions and $\mathcal{S}_t^{C,D} \in \mathcal{FAD}$ the feasible regions, for the controller state space when only the HoL delay objective is taken into account. If the PDR objective is considered, then $\mathcal{S}_t^{C,P} \in \mathcal{UFP}$ is the unfeasible state, whereas $\mathcal{S}_t^{C,P} \in \mathcal{FAP}$ represents the feasible state region. When the DP CMOO is performed, the feasibility of the newest state is given by intersecting the both regions such that $\{\mathcal{FADP}\} = \{\mathcal{FAD}\} \bigcap \{\mathcal{FAP}\}$ and the unfeasibility is denoted by the reunion of both unfeasible zones: $\{\mathcal{UFDP}\} = \{\mathcal{UFD}\} \bigcup \{\mathcal{UFP}\}$. In

other words, the controller state $\mathcal{S}_t^{C,DP} \in \mathcal{FADP}$ is feasible if and only if all active users are satisfied from the viewpoint of HoL packet delay and PDR constraints. Otherwise, the scheduler is unfeasible. This reasoning is shown in Eq. 7.4.

$$\mathcal{S}_t^{C,DP} \in \begin{cases} \{\mathcal{FADP}\}, \ if \ d_i^{HOL}[t] \le d_{i,L}^{\overline{HoL}}[t] \ and \ R_i^{PL}[t] \le R_i^{\overline{PL}}[t], \forall i \in \mathcal{U}_t \\ \{\mathcal{UFDP}\}, \ if \ \exists d_i^{HOL}[t] > d_{i,L}^{\overline{HoL}}[t] \ or \ \exists R_i^{PL}[t] > R_i^{\overline{PL}}[t], \forall i \in \mathcal{U}_t \end{cases} \tag{7.4}$$

The elements of the controller state space $\mathcal{S}_t^{C,DP}$ take into account the aggregate channel indicators and the special information about the queue states, arrival rates, PDR and HoL packet delay satisfaction. The state space for the DSR-CMOO MDP problems can then be defined by using Eq. 7.5:

$$\begin{aligned} \mathcal{S}_t^{C,DP} = \Big\{ & \mathcal{A}_{t-1}^{a,DP}, N_{CL}^{A}, \sigma_{CL}^{t}, ks_t, d_{ks}^{t}, \mu_{\overline{T}}^{t}, \sigma_{\overline{T}}^{t}, |\mathcal{U}_t|, & (a) \\ & \mu_D^t, \sigma_D^t, \mu_{\lambda D}^t, \sigma_{\lambda D}^t, N_{t,D}^{SAT}, N_{t,D}^{USAT}, d_i^{\overline{HoL}}, & (b) \\ & \mu_P^t, \sigma_P^t, \mu_{\lambda P}^t, \sigma_{\lambda P}^t, N_{t,P}^{SAT}, N_{t,P}^{USAT}, R_i^{\overline{PL}}, & (c) \\ & \mu_{qTX}^t, \sigma_{qTX}^t, \mu_{\overline{\lambda}}^t, \sigma_{\overline{\lambda}}^t, N_{t,Queues}^{Active}, N_{t,Queues}^{Unactive} \Big\} & (d) \end{aligned} \tag{7.5}$$

where the first part of this state (7.5.a) has already been defined in Chapter 6 with the amendment that $\mathcal{A}_{t-1}^{a,DP}$ is representing the action set focusing only on HoL delay requirement. Equation 7.5.b represents the HoL delay state elements where the set $\left(\mu_D^t, \sigma_D^t\right)$ represents the mean and STD HoL delay calculated based on Equations 4.6 and 4.7 from Chapter 4, $\left(\mu_{\lambda D}^t, \sigma_{\lambda D}^t\right)$ is the mean and STD values for the instantaneous difference $\lambda_{i,D}[t] = \left[d_{i,L}^{\overline{HoL}}[t] - d_i^{HoL}[t] \right]_+$ for each user $i \in \mathcal{U}_t$, and $N_{t,D}^{SAT}$ represents the normalized number of satisfied bearers from the HoL delay budget perspective at TTI *t*. Equation 7.5.c denotes the same parameter as Eq. 7.5.b being calculated for the packet drop rate for each active user. In Eq. 7.5.d, the data set $\left(N_{t,Queues}^{Active}, N_{t,Queues}^{Unactive}\right)$ represents the normalized number of active and un-active queues from the total number of active bearers at TTI *t*. When the delay requirement is very restrictive and a high percentage of active queues is

considered in the scheduling procedure, then the controller state $\mathcal{S}_t^{C,DP}$ feasibility is not reached based on Eq. 7.4 since a percentage of active bearers are considered to be in outage from the delay objective point of view. This aspect becomes more problematic when the DSR-CMOO problems focusing on the HoL delay and PDR constraints are solved with large PDR window T_w^{PDR} factors. Then, the feasibility is decided based on the reward function which can accept a small percentage of bearers to be in outage from the perspective of HoL delay and PDR objectives in order to model the DSR-CMOO MDP problems as episodic tasks.

7.2.3 Reward Function

The reward function which grants the DP DSR-CMOO MDP combinatorial problems considers the merged value of both HoL delay reward $\left(\mathcal{RW}_t^D\right)$ and PDR reward $\left(\mathcal{RW}_t^P\right)$ as shown in Eq. 7.6:

$$\begin{aligned}\mathcal{RW}_t^{DP}\left(\mathcal{A}_{t-1}^{a,DP},\mathcal{S}_{t-1}^{C,DP}\right)=\delta_D\cdot\mathcal{RW}_t^D\left(\mathcal{A}_{t-1}^{a,DP},\mathcal{S}_{t-1}^{C,DP}\right)\\ +\delta_P\cdot\mathcal{RW}_t^P\left(\mathcal{A}_{t-1}^{a,DP},\mathcal{S}_{t-1}^{C,DP}\right)\end{aligned} \tag{7.6}$$

where the set of $\left(\delta_D,\delta_P\right)\in\mathbb{R}_{[0,1]}$ represents the reward weights by respecting the property of $\delta_D+\delta_P=1$. When $\delta_D>\delta_P$, the learned scheduling policy focuses more on the HoL delay objective, whereas when $\delta_D<\delta_P$, the first priority of the exploited policy is constituted by the PDR objective.

The HoL delay and PDR reward functions can be determined similar to the GBR reward from Chapter 6 by considering the normalized instantaneous differences, the performance parameters $\left(d_i^{HoL},R_i^{PL}\right)$ and the 3GPP requirements. In particular, the relative HoL packet delay $\hat{\lambda}_i^D$ reported to the HoL delay requirement for each user and at each TTI can be calculated based on Eq. 7.7:

$$\hat{\lambda}_i^D\left[t\right]=\frac{\vartheta_D\cdot\bar{d}_i^{HoL}\left[t\right]-d_i^{HoL}\left[t\right]}{\vartheta_D\cdot\bar{d}_i^{HoL}},\ \forall i\in\mathcal{U}_t,\forall\vartheta_D\in\mathbb{R}_{[0,1]} \tag{7.7}$$

When $\hat{\lambda}_i^D[t] > 0$, the radio bearer $i \in \mathcal{U}_t$ is satisfied from the HoL delay perspective. Then, the HoL delay reward for each active radio bearer is calculated by using the following equation:

$$\mathcal{RW}_i^D[t] = \begin{cases} \hat{\lambda}_i^D[t], \ if \ \hat{\lambda}_i^D[t] \leq 0 \\ 1, \ if \ \hat{\lambda}_i^D[t] > 0 \end{cases} \tag{7.8}$$

The *intrinsic HoL delay reward* is obtained by summing the delay sub-rewards for each active bearer such that:

$$\mathcal{RWI}^D[t] = \sum_{i=1}^{|\mathcal{U}_t|} \mathcal{RW}_i^D[t] \tag{7.9}$$

Finally, the global delay reward can be calculated as a temporal difference between two consecutive intrinsic rewards as expressed in Eq. 7.10:

$$\mathcal{RW}^D[t] = \begin{cases} 1, & if \ \mathcal{RWI}^D[t] = 1 \\ \mathcal{RWI}^D[t] - \ell_D \cdot \mathcal{RWI}^D[t-1], & otherwise \end{cases} \tag{7.10}$$

where $\ell_D \in \{0,1\}$ decides whether the temporal intrinsic reward difference should be considered or not. For the PDR particular sub-reward, the normalized instantaneous difference $\hat{\lambda}_i^P$ can be computed online in the similar way as Eq. 7.7 being expressed as follows:

$$\hat{\lambda}_i^P[t] = \frac{\overline{R_i^{PL}}[t] - R_i^{PL}[t]}{\overline{R_i^{PL}}[t]}, \ \forall i \in \mathcal{U}_t \tag{7.11}$$

If $\hat{\lambda}_i^P[t] > 0$, then the PDR objective is satisfied for a given time window T_w^{PDR}. When $\hat{\lambda}_i^P[t] < 0$, then the PDR reward for each bearer $i \in \mathcal{U}_t$ takes the normalized value as a punishment, a fact which is exposed by Eq. 7.12 followed by the intrinsic PDR reward which is determined based on Eq. 7.13.

$$\mathcal{RW}_i^P[t] = \begin{cases} \hat{\lambda}_i^P[t], \ if \ \hat{\lambda}_i^P[t] \leq 0 \\ 1, \ if \ \hat{\lambda}_i^P[t] > 0 \end{cases} \tag{7.12}$$

$$\mathcal{RWI}^{P}\left[t\right]=\sum_{i=1}^{|\mathcal{U}|}\mathcal{RW}_{i}^{P}\left[t\right] \tag{7.13}$$

The global PDR reward is calculated in a similar way to other QoS objectives such as GBR and HoL delay:

$$\mathcal{RW}^{P}\left[t\right]=\begin{cases}1, & if\ \ \mathcal{RWI}^{P}\left[t\right]=1\\ \mathcal{RWI}^{P}\left[t\right]-\ell_{P}\cdot\mathcal{RWI}^{P}\left[t-1\right], & otherwise\end{cases} \tag{7.14}$$

where $\ell_{P}\in\{0,1\}$ determines the type of the PDR reward. In the DSR-CMOO MDP problems, the parameters (ℓ_{D},ℓ_{P}) are very important since the performance of scheduling policies can be improved when a proper setting of these factors is achieved. For the current purpose of the DP optimization problems, the settings of $(\ell_{D}=1,\ell_{P}=1)$ are considered, which means that both objective rewards consider the temporal difference between two consecutive intrinsic rewards. For very large PDR time windows, the intrinsic temporal difference becomes mandatory in the overall reward computation since it can detect any difference between consecutive TTIs which can appear when the long term PDR observations are computed.

For the HoL delay reward, the requirement fraction factor ϑ_{D} is very important since the episodic state nature is directly connected to this parameter and to the number of active bearers. For instance, when a large number of active queues has to be satisfied at each TTI, then even if the requirement fraction is $\vartheta_{D}=1$, the DP DSR-CMOO episodic tasks are not guaranteed. In order to avoid this drawback, the HoL packet delay reward suffers the following modification:

$$\mathcal{RW}^{D}\left[t\right]=\begin{cases}1, & if\ \ \mathcal{RWI}^{D}\left[t\right]\geq\kappa_{D}\\ \mathcal{RWI}^{D}\left[t\right]-\ell_{D}\cdot\mathcal{RWI}^{D}\left[t-1\right], & if\ \ \mathcal{RWI}^{D}\left[t\right]<\kappa_{D}\end{cases} \tag{7.15}$$

where $\kappa_{D}\in\mathbb{R}_{[0,1]}$ is the relaxation parameter which permits to increase the number of terminal states when the scheduler reward is $\mathcal{RW}_{t}^{DP}=1$ during the exploration period. In this sense, the terminal state can contain some active bearers which are considered to be in outage from the HoL delay point of view $\left(N_{t,D}^{USAT}\geq 0\right)$. In this section, the delay reward from Eq. 7.15 is used in order to

detect the scheduling rules that aim to reduce the mean of HoL delays with higher standard deviation values. The scheduling policies should be able to combine the scheduling rules with high or low delay variations in order to increase the number of TTIs when all the bearers are satisfied from the lower HoL delay requirement point of view. In the following, the performances of the obtained scheduling policies are discussed for the CBR and VBR traffic types.

7.2.4 Performance Evaluation of Sustainable Scheduling Policies Focusing on HoL Packet Delay and PDR Objectives

The performance of the proposed scheduling policies is evaluated based on the mean percentage of TTIs for different DP multi-objective satisfaction levels and based on the mean percentage of TTIs for different reward types. The rest of this sub-section is organized as follows: Sub-section 7.2.4.1 presents the simulation scenario, Sub-section 7.2.4.2 highlights the performance of sustainable scheduling policies for the CBR traffic type and finally, Sub-section 7.2.4.3 addresses the simulation results of DP policies by using the VBR traffic type.

7.2.4.1 Simulation Scenario

The simulation scenario which is used to solve the DSR-CMOO MDP problem from Eq. 7.1 uses the parameters shown in Table 6.3 and the settings of parameters from Table 6.4 for the RL algorithms. The HoL delay requirements are switched at each 1000 TTIs by using the constraints which can be found in Table 2.1 from Chapter 2, such as $d_i^{\overline{HoL}} \in \{50,100,150,200,250,300\}\, ms$. When the HoL delay exceeds any of the exposed requirements, then the packet is dropped and declared lost. The number of active bearers is changed in the same intervals of time in the domain of $|\mathcal{U}_t| \in [15;120]$. In this sense, the delay fraction parameter is $\vartheta_D = 0.1$ and the considered relaxation parameter is $\kappa_D = 0.9$ in order to increase the possibilities of reaching the terminal states when higher traffic load is

considered for more restrictive HoL delay requirements. The DP multi-objective reward function (Eq. 7.6) considers equal weights for the HoL delay and PDR rewards $\left(\delta_D = \delta_P = 0.5\right)$ which gives in fact the same level of importance for both objectives which are involved in the learned scheduling policies.

For the PDR objective, the requirements are changed at each 1000 TTIs by using the following 3GPP parameters $\bar{R}_i^{PL} \in \left\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\right\}$ being exposed in Table 2.1 from Chapter 2. The time window length which is used for the online PDR computation is considered to be similar to the median filter length, and in this particular case, different MLPNN functions are trained by using different parameterizations for the windowing factor.

Another important factor which concerns the filter time window $T_w^{PDR} = T_w^M$ is the maximum number of users which can be scheduled at each TTI $\left(N_{Sched}^{Max}\right)$. Through certain special simulations which are not considered in the current study, it can be concluded that for a maximum number of users of $\left|\mathcal{U}_t\right| = 120$, the limit on the maximum number of schedulable users N_{Sched}^{Max} does not have a great influence on the quality of the scheduling policies. In this case, the number of users which are scheduled at each TTI varies in the interval of [6, 12], which does not affect the percentage of satisfied bearers when the maximum limit of schedulable users $N_{Sched}^{Max} = 10$ is considered. For a higher traffic load (e.g. hundreds of VoIP bearers), the maximum number of schedulable users becomes crucial for the HoL delay objective and the windowing factor has to be modified accordingly. The simulation results are conducted through CBR and VBR traffic types which are generated by following the parameters from Table 6.3. The performances of the scheduling policies are measured in terms of:

1. The mean percentage of TTIs when the active bearers are satisfied in a percentage of x% $\left(\bar{p}_{TTI}^{DP,x\%}\right)$ for the multi-objective evaluation. Also, in Appendix H, the mean percentages of TTIs in terms of particular objectives $\left(\bar{p}_{TTI}^{D,x\%}, \bar{p}_{TTI}^{P,x\%}\right)$ are analyzed for different settings of the PDR

median filter. The STD values of the mean percentage of TTIs for different DP performance levels play a crucial role in determining the sustainability of the proposed scheduling policies.

2. The mean percentage of TTIs when the testing rewards are punishments, moderate and maximized $\left(\overline{p}_{TTI}^{DP,PSH}, \overline{p}_{TTI}^{DP,mRW}, \overline{p}_{TTI}^{DP,MRW}\right)$ and the associated STD error values denote the capability of each policy of recovering the unfeasible states. Appendix H considers the impact of particular objectives in the mean percentages of TTIs for different reward types such as: $\left(\overline{p}_{TTI}^{D,PSH}, \overline{p}_{TTI}^{D,mRW}, \overline{p}_{TTI}^{D,MRW}, \overline{p}_{TTI}^{P,PSH}, \overline{p}_{TTI}^{P,mRW}, \overline{p}_{TTI}^{P,MRW}\right)$.

The idea is to follow the discrepancy between $\overline{p}_{TTI}^{DP,100\%}$ and $\overline{p}_{TTI}^{DP,MRW}$ in order to detect the characteristics of the scheduling rules and to decide the best scheduling policy from the perspective of the aforementioned indicators.

7.2.4.2 DSR-CMOO MDP Focusing on HoL Packet Delay and PDR Objectives with the CBR Traffic Type

The performances of the scheduling policies for the DP DSR-CMOO optimization problems are analyzed and compared by using different windowing factors when the rate of the dropped packets is calculated. For the CBR traffic type, the arrival rates fluctuate at each 1000 TTIs in the domain of the following elements: $\overline{T} = \{32;64;128;256;512;1024\}\ kbps$. For restrictive PDR time windows when $T_w^{PDR} \in [8,66]\ TTIs$ and $\rho = 5.5$, the QVMAX2 and ACLA scheduling policies outperform other techniques from the viewpoint of the combined delay and PDR objectives. As shown in Fig. 7.1, the mean percentage of TTIs for 100% DP satisfied bearers is nearly $\overline{p}_{TTI}^{DP,100\%} = 65\%$, which indicates a gain in number of TTIs of about 15% when compared with the GPF-LOG rule or the QV2 scheduling policy. The worst performance is obtained when the QV policy is exploited by indicating a percentage of $\overline{p}_{TTI}^{DP,100\%} = 9\%$ which is comparable with the performance obtained when the GPF-EXP1 scheduling rule is performed.

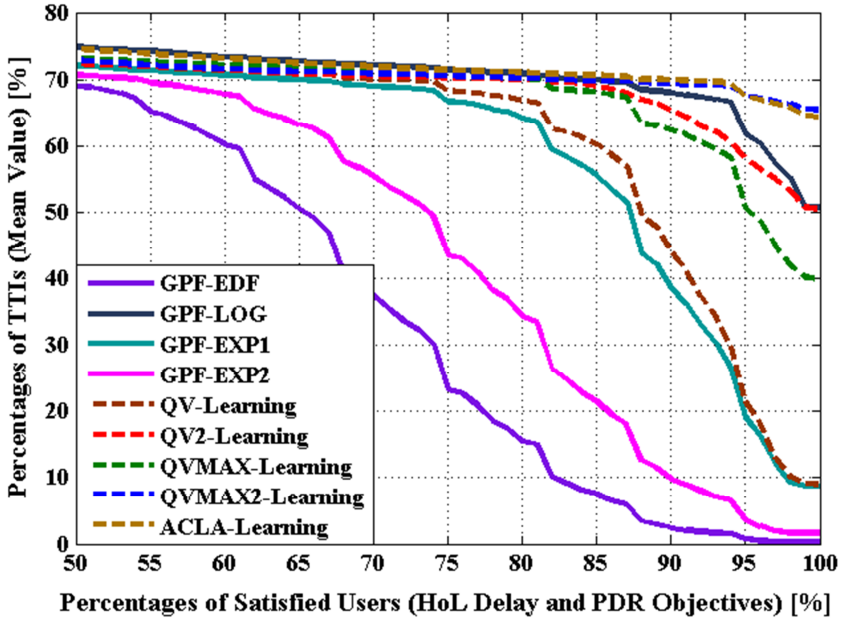


Fig. 7.1 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the CBR Traffic Type and the Windowing Factor of $\rho = 5.5$

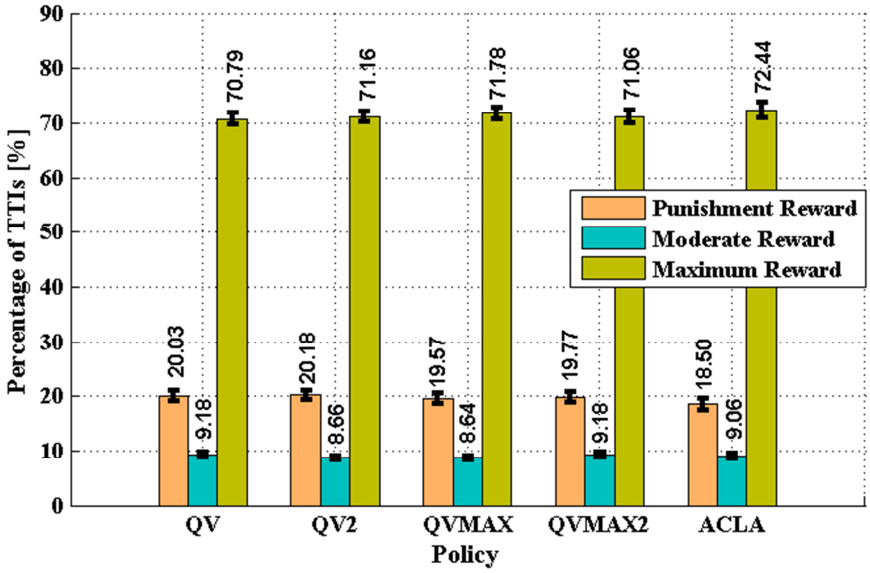


Fig. 7.2 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the CBR Traffic Type and the Windowing Factor of $\rho = 5.5$

The mean percentage of TTIs for the reward types from Fig. 7.2 indicates a percentage of $\overline{p}_{TTI}^{DP,MRW} > 70\%$ which involves in fact the idea that the number of episodes is greater than the number of states when $d_i^{HoL} < 0.1 \cdot \overline{d}_i^{HoL}, \forall i \in \mathcal{U}_t$. The scheduling policies obtained based on the ACLA and QVMAX2 RL algorithms

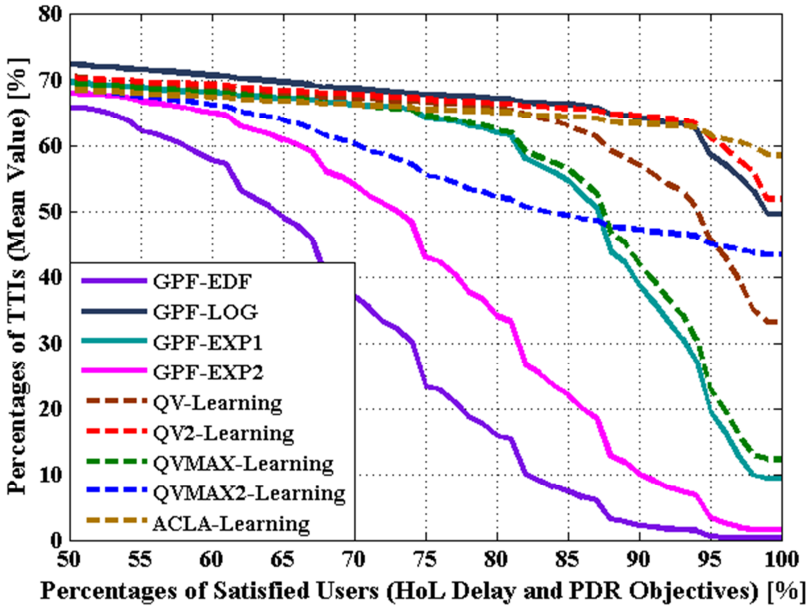


Fig. 7.3 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the CBR Traffic Type and the Windowing Factor of $\rho = 50$

provide the lowest discrepancy between $\bar{p}_{TTI}^{DP,100\%} \simeq 65\%$ and $\bar{p}_{TTI}^{DP,MRW} = 72.44\%$ by indicating at the same time, that the provided policies are able to minimize the standard deviation of HoL delays. The QV scheduling policy procedure follows the trajectory imposed by the GPF-EXP1 scheduling rule which in fact indicates a performance of $\bar{p}_{TTI}^{DP,100\%} = 9\%$ when in reality the mean percentage of TTIs with maximum rewards is $\bar{p}_{TTI}^{DP,MRW} = 70.79\%$. This way, it can be concluded that by applying the GPF-EXP1 scheduling rule to a large number of TTIs, the scheduling policy provides higher deviations for the HoL delay of each active bearer.

When larger windowing factors are considered (e.g., $\rho = 50$ in Fig.7.3), the overall scheduling performance is affected by the fact that the PDR objective should be satisfied for a larger time window domain such that $T_w^{PDR} \in [75, 600]$, which decreases the performance of $\bar{p}_{TTI}^{DP,100\%}$ percentages, as shown in Fig. 7.3. The best performance is obtained when the ACLA policy is exploited by indicating the percentage of $\bar{p}_{TTI}^{DP,100\%} = 59\%$ when all active bearers are satisfied from the perspective of HOL delay and PDR objectives. When the reward performances are analyzed in Fig. 7.4, the percentage of TTIs for the maximum

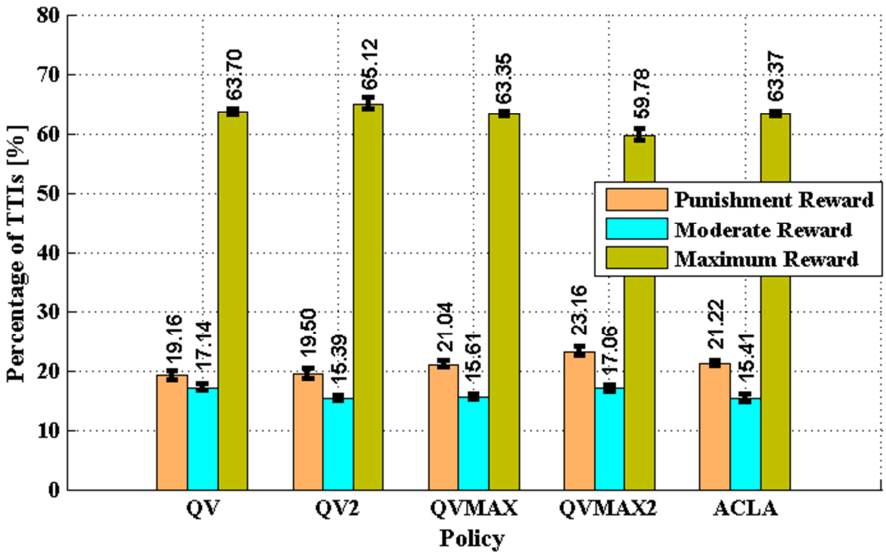


Fig. 7.4 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the CBR Traffic Type and the Windowing Factor of $\rho = 50$

reward is $\overline{p}_{TTI}^{DP,MRW} > 59\%$ for the obtained scheduling policies. In particular, the ACLA learning procedure is able to learn better the optimal DP policies due to the actor-critic scheme characteristic and to the fact that the temporal difference between two consecutive intrinsic rewards is considered in the DP multi-objective reward function computation. The performance difference between $\overline{p}_{TTI}^{DP,100\%}$ and $\overline{p}_{TTI}^{DP,MRW}$ is more than 50% for the QVMAX scheduling policy when the GPF-EXP1 scheduling discipline is applied for a large number of TTIs.

Figure 7.5 indicates the scheduling performance for the DP DSR-CMOO MDP problems when the time window varies in the interval of $T_w^{PDR} \in [150,1200]$ number of TTIs when the considered static windowing factor is $\rho = 100$. The overall performance of the mean percentage of DP feasible TTIs $\overline{p}_{TTI}^{DP,100\%}$ indicates a loss of 10% when compared with the scenario of $\rho = 5.5$. The QV2, QVMAX, QVMAX2 and ACLA learning procedures offer close performances from the percentages of TTIs when all active bearers are 100% satisfied from the viewpoint of HoL delay and PDR objectives. The gain of $\overline{p}_{TTI}^{DP,100\%}$ percentage obtained by ACLA is about 10% when compared with GPF-LOG and about 44% when compared with the QV policy or with the GPF-EXP1 scheduling rule.

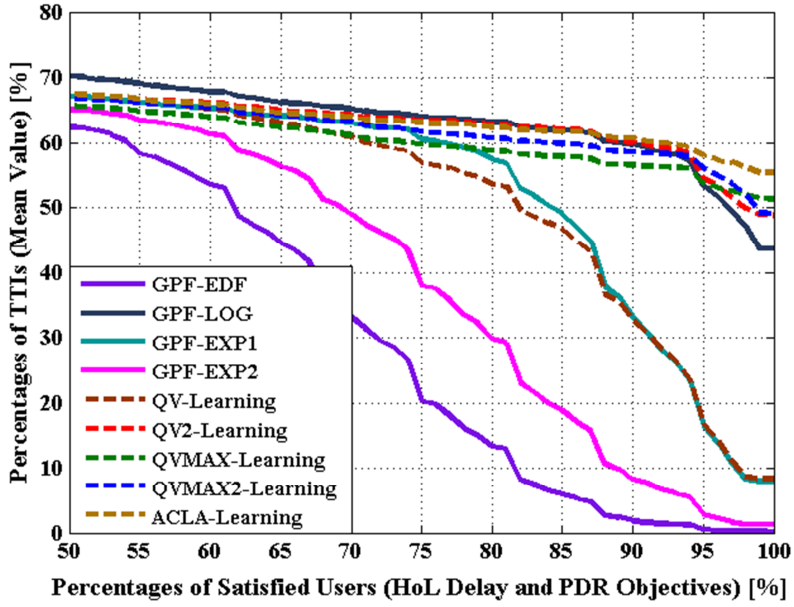


Fig. 7.5 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the CBR Traffic Type and the Windowing Factor of $\rho = 100$

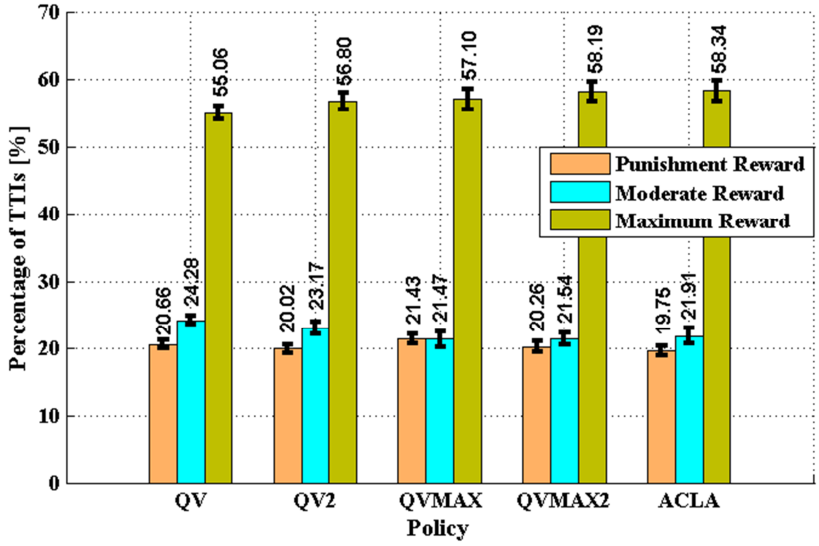


Fig. 7.6 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the CBR Traffic Type and the Windowing Factor of $\rho = 100$

The impact of selecting the GPF-EXP1 scheduling rule by the QV policy is highlighted in Fig. 7.6 which illustrates a percentage of $\overline{p}_{TTI}^{DP,MRW} = 55.06\%$ when the percentage of TTIs with 100% DP satisfied bearers is about $\overline{p}_{TTI}^{DP,100\%} \approx 9\%$. By applying the GPF-EXP1, GPF-EXP2 and GPF-EDF scheduling rules only when the moderate rewards are considerable higher (when compared

with the moderate reward amount received for the GPF-LOG rule), then ACLA scheduling policy is able to outperform other scheduling techniques from the viewpoint of the mean percentage $\overline{p}_{TTI}^{DP,100\%}$ of DP feasible TTIs.

The evolution on rewards for the ACLA and QVMAX2 exploited policies when the windowing factors are $\rho \in \{5.5; 50; 100; 200; 300; 400; 500\}$ is illustrated in Fig. 7.7. It is important to notice that the windowing factor variability does not have any impact on the particular reward which is focused only on the HoL delay objective but it affects the multi-objective reward $\mathcal{RW}_t^{DP}$ at each TTI. When the windowing factor is $\rho = 500$, more than 50% of TTIs with maximum rewards are declared lost when compared with the scheduling policy with the windowing factor of $\rho = 5.5$. On the other hand, the percentage of TTIs with a moderate reward registers an increase of 50% when compared with the case of $\rho = 5.5$. By increasing the windowing factor for both ACLA and QVMAX2 policies, the percentage of TTIs with maximum rewards is decreased in the detriment of the percentage of TTIs with moderate rewards while maintaining the percentage of punishments constant for the considered domain of windowing factors.

Figure 7.8 analyses the performance of the RL scheduling policies, GPF-LOG and GPF-EXP2 scheduling rules from the perspective of the scheduling quality indicators of $\left\{\overline{p}_{TTI}^{DP,80\%}, \overline{p}_{TTI}^{DP,85\%}, \overline{p}_{TTI}^{DP,90\%}, \overline{p}_{TTI}^{DP,94\%}, \overline{p}_{TTI}^{DP,96\%}, \overline{p}_{TTI}^{DP,100\%}\right\}$ when the windowing factor takes discrete values from $\rho = 5.5$ to $\rho = 500$. From the $\overline{p}_{TTI}^{DP,100\%}$ percentage point of view, the RL policies outperform GPF-LOG and GPF-EXP2 scheduling rules for each considered windowing factor discrete value. When $\rho = 5.5$, the ACLA policy performs better than GPF-LOG for the entire domain of DP bearer satisfaction. When the windowing factor increases, the long term PDR affects the global DP reward function and the controller is not able to take optimal actions at each TTI. For these reasons, ACLA is not able to outperform GPF-LOG from the viewpoint of $\left\{\overline{p}_{TTI}^{DP,80\%}, \overline{p}_{TTI}^{DP,85\%}\right\}$ percentages when the windowing factor is $\rho > 200$. However, the ACLA policy indicates a set of

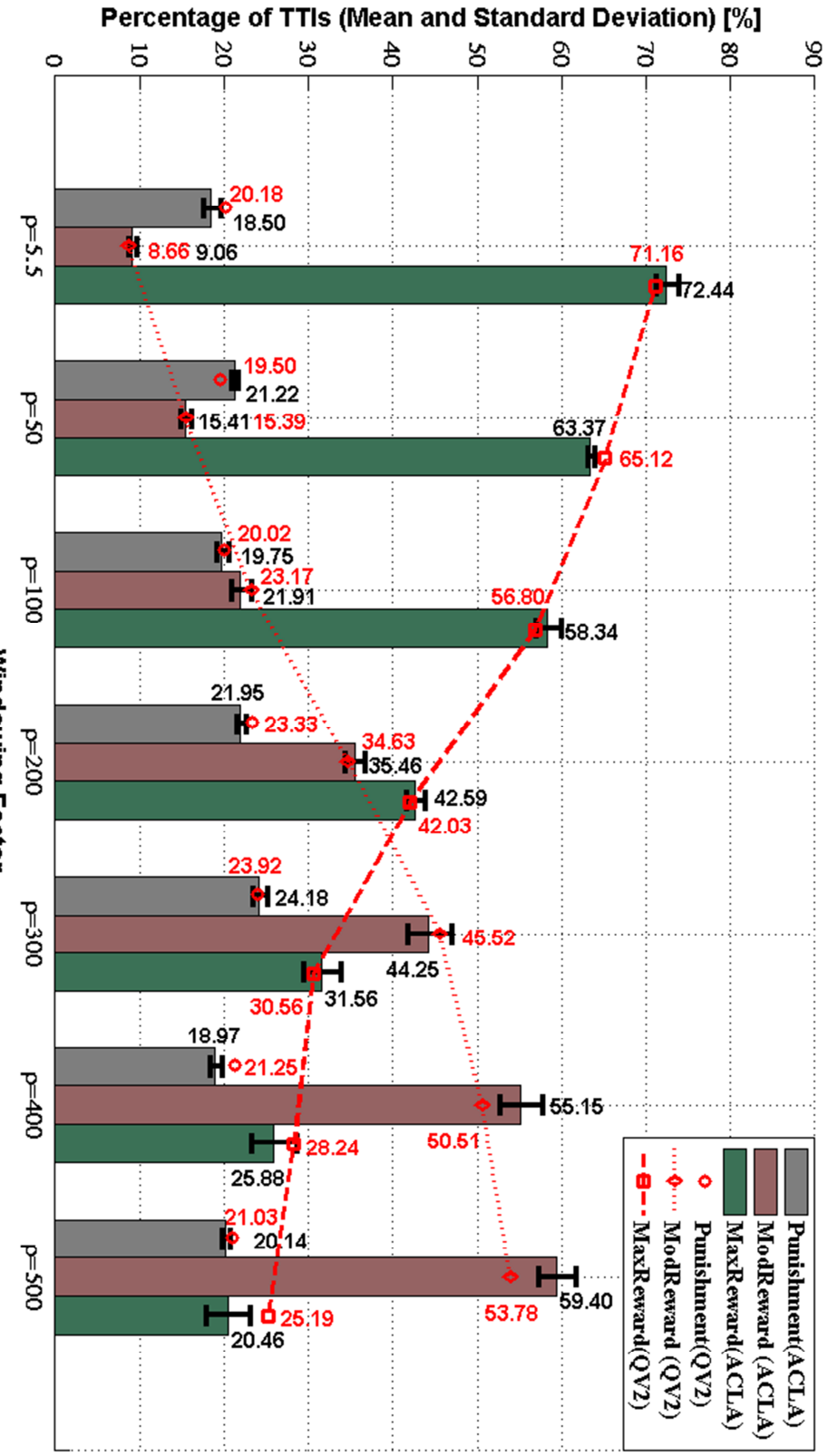


Fig. 7.7 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the CBR Traffic Type and Variable Windowing Factors

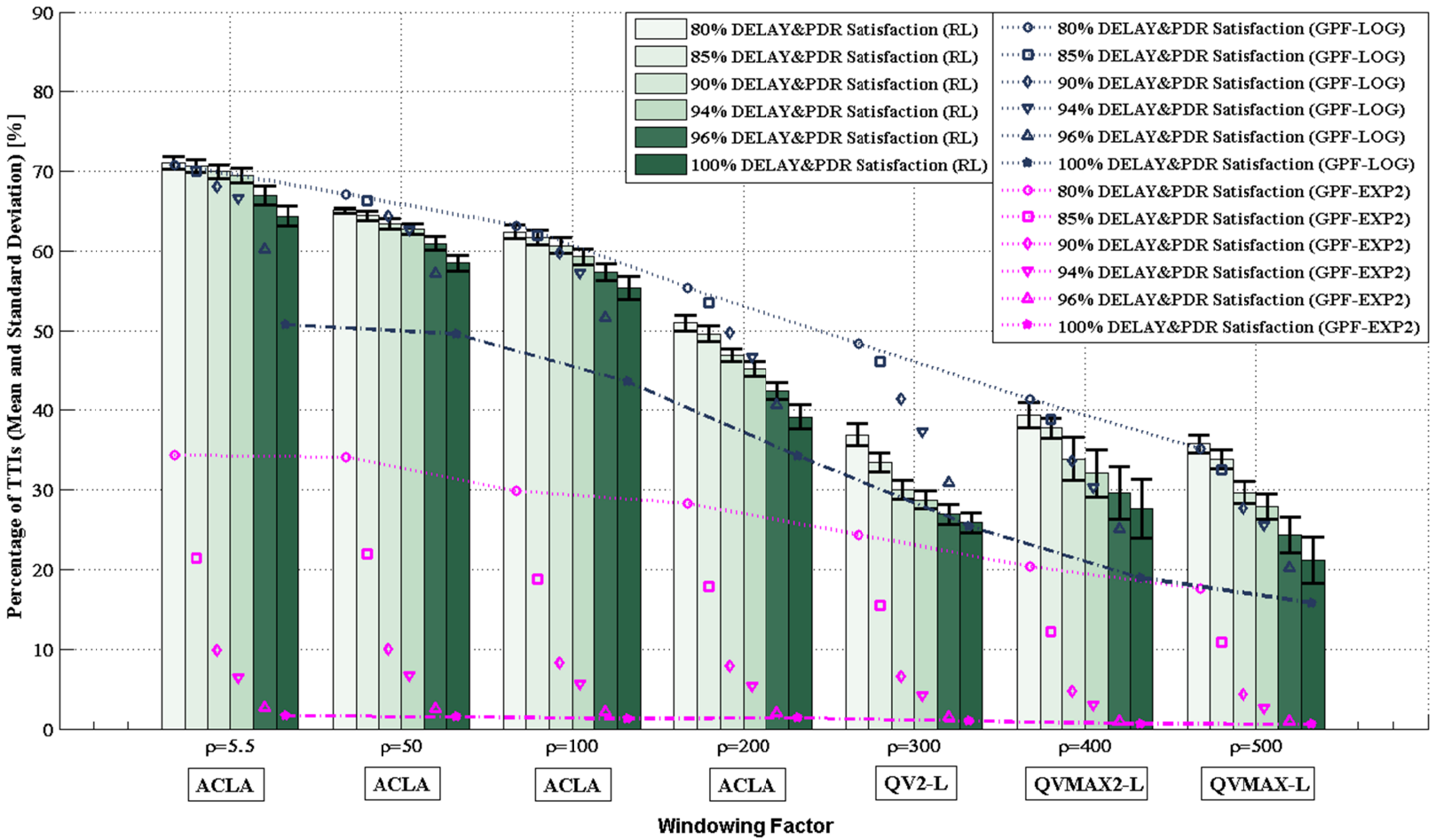


Fig. 7.8 Mean Percentages of TTIs for $\{100\%;96\%;94\%;90\%;85\%;80\%\}$ DP Satisfaction with the CBR Traffic Type and Variable Windowing Factors

the gains of about $\{3\%;6.6\%;13.6\%\}$ for $\left\{\bar{p}_{TTI}^{DP,94\%}, \bar{p}_{TTI}^{DP,96\%}, \bar{p}_{TTI}^{DP,100\%}\right\}$ when the windowing factor is $\rho = 5.5$, and when this factor becomes equal to $\rho = 500$, the QVMAX policy outperforms GPF-LOG and GPF-EXP2 by indicating the gain set of $\{2.4\%;3.5\%;5.9\%\}$ for the percentage set of $\left\{\bar{p}_{TTI}^{DP,94\%}, \bar{p}_{TTI}^{DP,96\%}, \bar{p}_{TTI}^{DP,100\%}\right\}$. The rest of the simulation results for the CBR traffic type are listed in Appendix H.

According to Figure 7.8, the scheduling policies obtained by using different RL approaches for different windowing factor settings are able to outperform other proposals and static scheduling rules when the performance criterion of the mean percentage of DP feasible TTIs $\bar{p}_{TTI}^{DP,100\%}$ is considered. Based on the simulation results provided in Sub-section H.3, the same policies minimize the number of punishments and the associated STD values. For these reasons, the obtained policies are considered sustainable in the long term LTE scheduling.

7.2.4.3 DSR-CMOO MDP Focusing on HoL Packet Delay and PDR Objectives with the VBR Traffic Type

When the VBR traffic type is simulated under the DP DSR-CMOO MDP problems, the same indicators are studied in order to highlight the characteristics of the scheduling rules and the performance of the learned policies. For instance, in Fig. 7.9, the QV policy outperforms other candidates in the interval of $\left[\bar{p}_{TTI}^{DP,95\%}, \bar{p}_{TTI}^{DP,100\%}\right]$. When the performance interval of $\left[\bar{p}_{TTI}^{DP,20\%}, \bar{p}_{TTI}^{DP,95\%}\right]$ is considered, the ACLA policy becomes the best choice since it follows the GPF-LOG rule TTI-by-TTI. The QV policies sacrifice the performance domain of $\left[\bar{p}_{TTI}^{DP,20\%}, \bar{p}_{TTI}^{DP,95\%}\right]$ in order to increase the mean percentage of TTIs when the active bearers are 100% satisfied from the viewpoint of HoL delay and PDR objectives. The performance discrepancy between $\bar{p}_{TTI}^{DP,100\%}$ and $\bar{p}_{TTI}^{DP,MRW}$ is indicated in Fig. 7.10 for all the analyzed RL algorithms when the windowing factor is $\rho = 5.5$. ACLA indicates the highest percentage of TTIs with

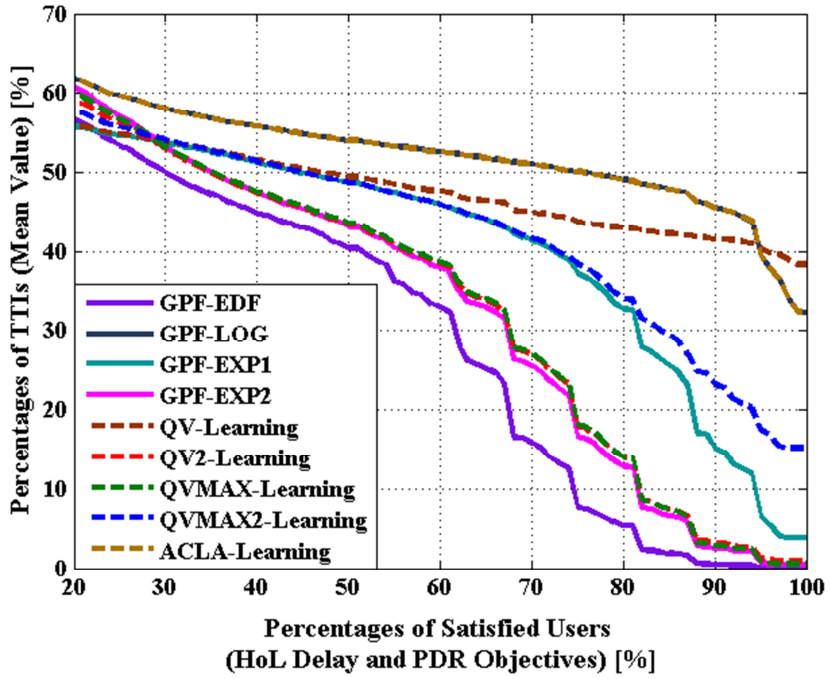


Fig. 7.9 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the VBR Traffic Type and the Windowing Factor of $\rho = 5.5$

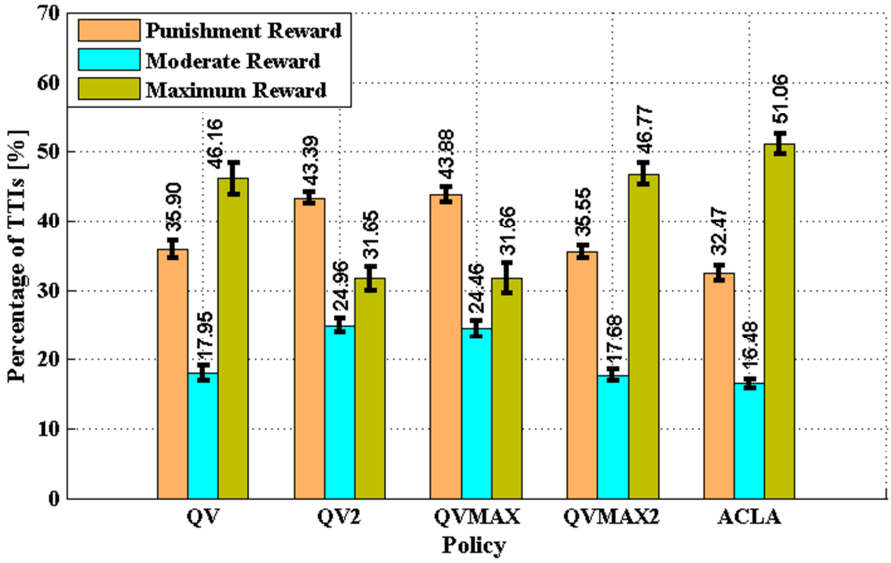


Fig. 7.10 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the VBR Traffic Type and the Windowing Factor of $\rho = 5.5$

maximum rewards $\bar{p}_{TTI}^{DP,MRW}$ in the exploitation period whereas the highest mean percentages of punishment $\bar{p}_{TTI}^{DP,PSH}$ and moderate $\bar{p}_{TTI}^{DP,mRW}$ rewards are obtained when the QV2 or QVMAX policies are exploited. When larger windowing factors

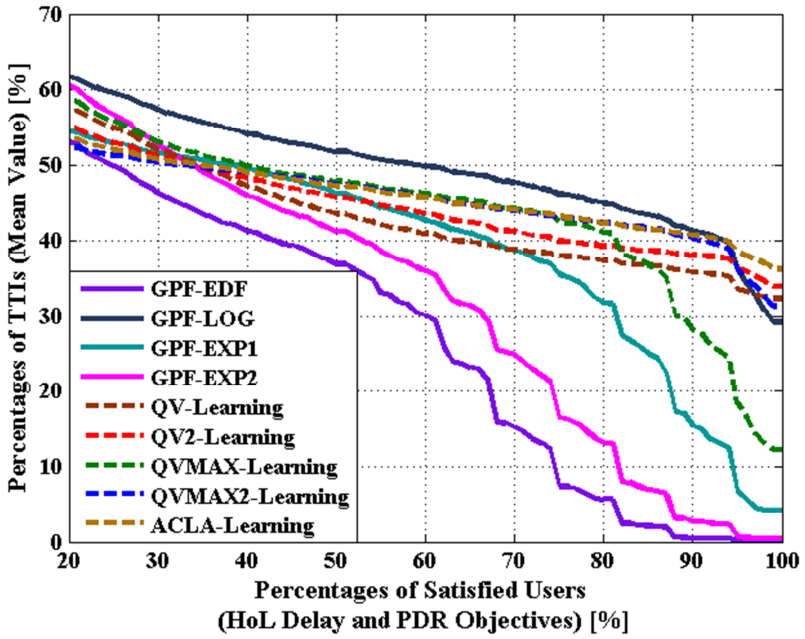


Fig. 7.11 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the VBR Traffic Type and the Windowing Factor of $\rho = 100$

are considered in the DP DSR-CMOO optimization problems such as $\rho = 100$, the DSR-CMOO problem remains focused on maximizing the percentage of DP feasible TTIs $\bar{p}_{TTI}^{DP,100\%}$ but affects the mean percentage of TTIs lower than $\bar{p}_{TTI}^{DP,75\%}$ due to the fact that for a large number of active bearers, the controller is not able to take optimal actions which can increase the number of moderate rewards. In Fig. 7.11, all RL algorithms except QVMAX outperform the main candidate GPF-LOG scheduling rule by indicating a gain from the viewpoint of the mean percentage of DP feasible TTIs $\bar{p}_{TTI}^{DP,100\%}$ of about [2-7]%. For the performance interval of $\left[\bar{p}_{TTI}^{DP,20\%}, \bar{p}_{TTI}^{DP,95\%}\right]$, GPF-LOG performs better than any other RL candidates. The reason for losing the scheduling policy optimality when $\rho = 100$ is highlighted in Fig. 7.12 where the percentage of punishments is almost equal with the percentage of maximum rewards for all RL candidates. The mean percentage of TTIs with maximum rewards is reduced by about 10% when compared against the case of $\rho = 5.5$ which in fact is taken by the percentage of TTIs with punishment rewards. When the windowing factor is $\rho = 300$, the maximum time window for 120 active bearers is 3600 TTIs. In this case, the

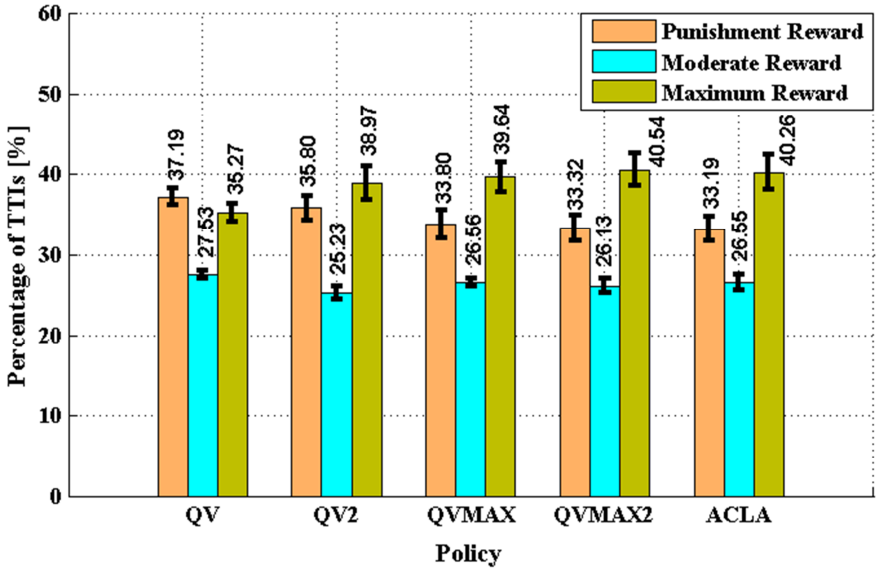


Fig. 7.12 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the VBR Traffic Type and the Windowing Factor of $\rho = 100$

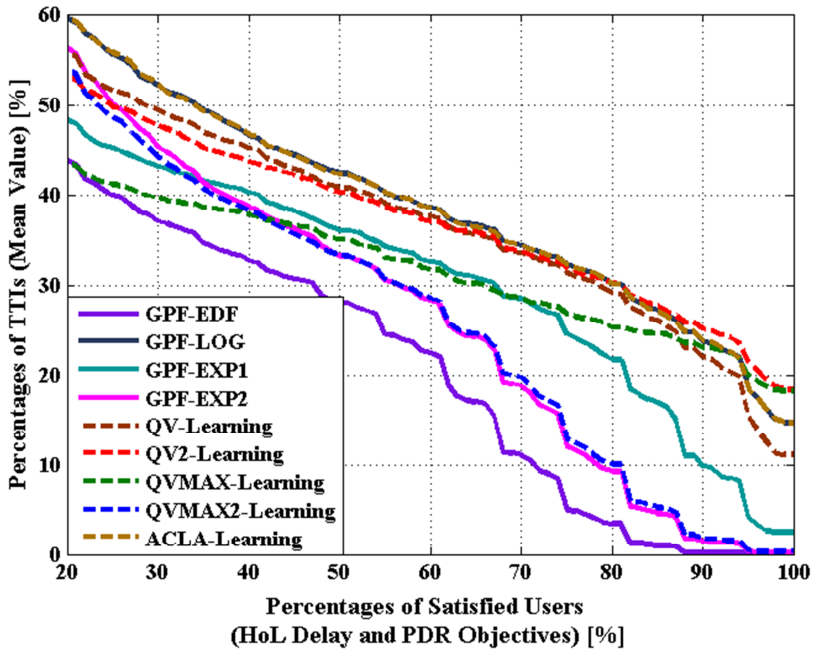


Fig. 7.13 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the VBR Traffic Type and the Windowing Factor of $\rho = 300$

ACLA learning selects the GPF-LOG scheduling rule for almost the entire exploitation session while QVMAX and QV2 scheduling policies increase the mean percentage $\overline{p}_{TTI}^{DP,100\%}$ of about 4% when compared with ACLA or GPF- LOG. In fact, the QV2 policy performs better than other candidates when considering

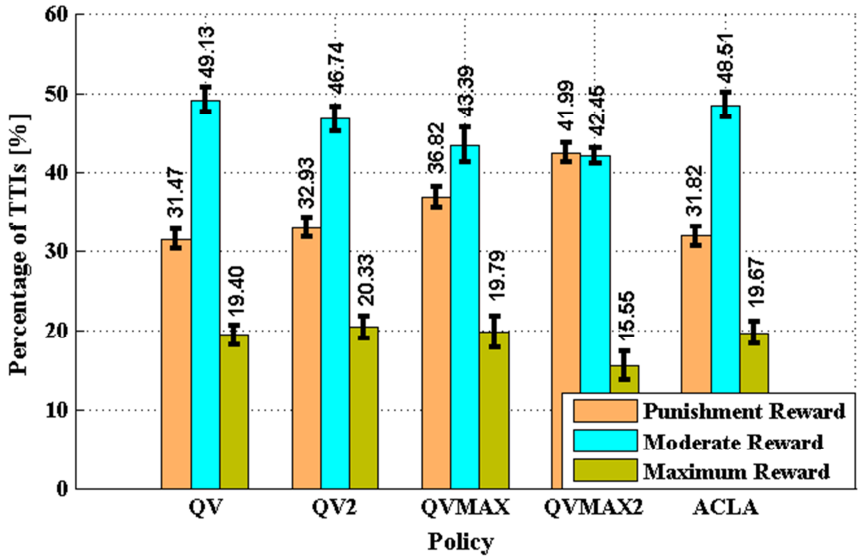


Fig. 7.14 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the VBR Traffic Type and the Windowing Factor of $\rho = 300$

the performance interval of $\left[\bar{p}_{TTI}^{DP,85\%}, \bar{p}_{TTI}^{DP,100\%}\right]$. The number of maximum rewards is seriously degraded when compared with other windowing factor cases for all RL approaches. The worst option is represented by the QVMAX2 policy (Fig. 7.14) which indicates an equal amount of punishment and moderate rewards. On the other hand, the highest amount of moderate rewards is obtained through QV and ACLA algorithms which provide a close performance for a wider DP performance domain when compared with any other candidates.

The evaluation on the reward type based on the windowing factor for ACLA and QVMAX2 policies is highlighted in Fig. 5.15. The percentage of TTIs with punishment rewards is higher, when compared with the CBR case, by about 10% for the considered domain of windowing factors. The same behavior of $\bar{p}_{TTI}^{DP,mRW}$ and $\bar{p}_{TTI}^{DP,MRW}$ is registered for both ACLA and QVMAX2 algorithms when the windowing factor takes values from $\rho = 5.5$ to $\rho = 500$. By increasing the time window for the PDR performance evaluation for ACLA and QVMAX2 learned policies, the mean percentage of TTIs with maximum rewards $\bar{p}_{TTI}^{DP,MRW}$ is decreased with more than 40% and $\bar{p}_{TTI}^{DP,mRW}$ is increased with the same amount for $\rho = 500$ when compared with the more restrictive time domain of $\rho = 5.5$.

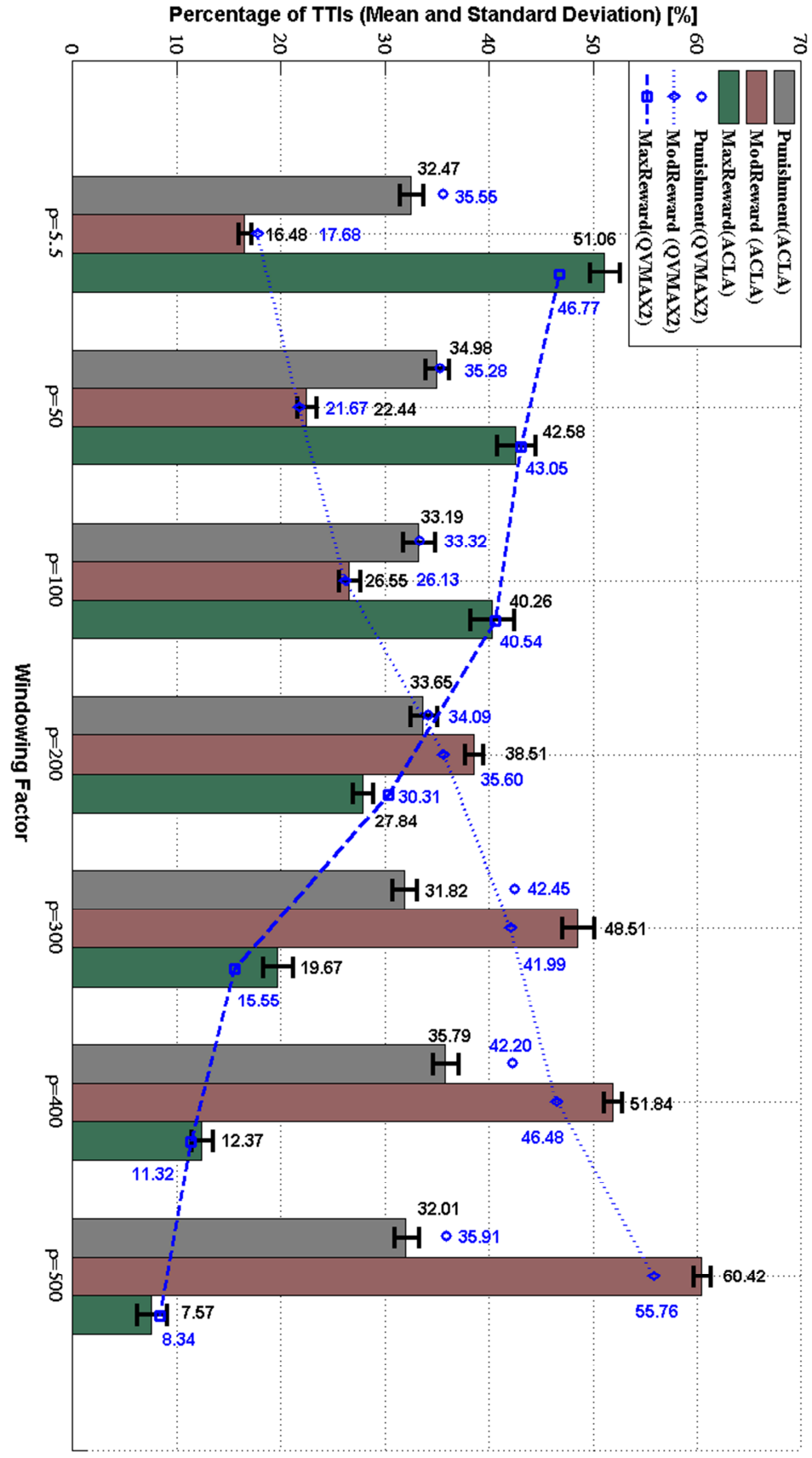


Fig. 7.15 Mean Percentages of TTIs for Punishment, Moderate and Maximum DP Rewards with the VBR Traffic Type and Variable Windowing Factors

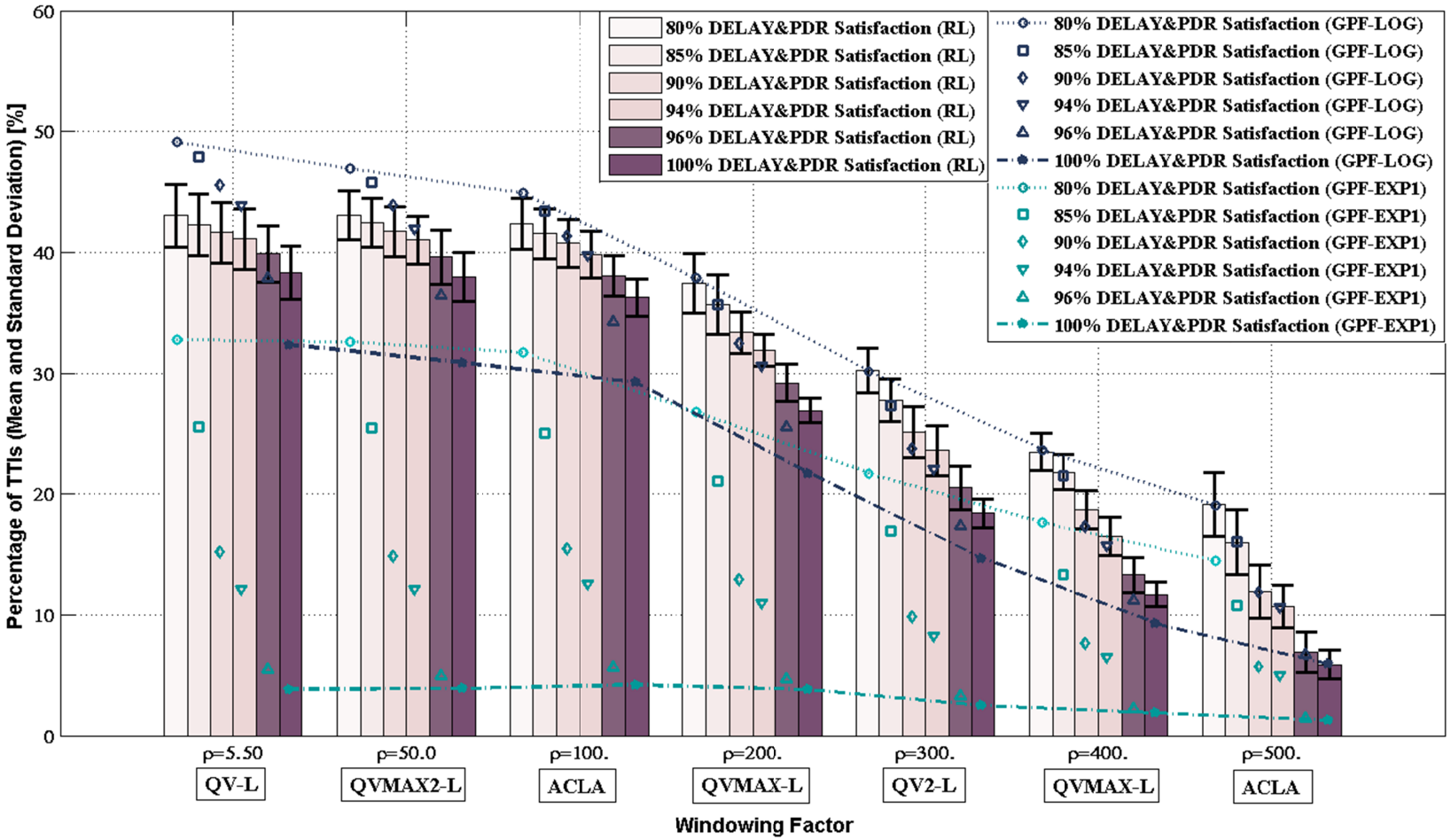


Fig. 7.16 Mean Percentages of TTIs for $\{100\%;96\%;94\%;90\%;85\%;80\%\}$ DP Satisfaction with the VBR Traffic Type and Variable Windowing Factors

Figure 7.16 shows the best performance of the proposed set of scheduling policies from the viewpoint of $\left\{\bar{p}_{TTI}^{DP,80\%}, \bar{p}_{TTI}^{DP,85\%}, \bar{p}_{TTI}^{DP,90\%}, \bar{p}_{TTI}^{DP,94\%}, \bar{p}_{TTI}^{DP,96\%}, \bar{p}_{TTI}^{DP,100\%}\right\}$ performance indicators. The learned scheduling policies outperform the classical scheduling rules of GPF-LOG and GPF-EXP1 when the percentages of $\left\{\bar{p}_{TTI}^{DP,96\%}, \bar{p}_{TTI}^{DP,100\%}\right\}$ are taken into account for the considered domain of windowing factors. It is obvious that for large windowing factors such as $\rho = 400$ and $\rho = 500$, both QVMAX and ACLA policies follow the GPF-LOG scheduling rule for the entire performance domain of $\left[\bar{p}_{TTI}^{DP,80\%}, \bar{p}_{TTI}^{DP,100\%}\right]$.

By using the DSR-CMOO MDP problem proposed in Eq. 7.1 for both CBR and VBR traffic types, the proposed scheduling policies are able to increase the number of feasible TTIs from the viewpoint of HoL delay and PDR objectives only for the performance domain of $\left[\bar{p}_{TTI}^{DP,95\%}, \bar{p}_{TTI}^{DP,100\%}\right]$. Due to the main fact that the mean percentage $\bar{p}_{TTI}^{DP,MRW}$ is greater than $\bar{p}_{TTI}^{DP,100\%}$ for all RL approaches, the scheduling policies are not able to take optimal actions in order to satisfy lower percentages of bearers and instead are focused for the entire exploitation period on maximizing the percentage of TTIs when the bearers are 100% satisfied from the viewpoint of HoL packet delay and PDR multi-objective criterion.

In Appendix H, the percentages of TTIs (mean and STD) for different levels of objective satisfaction such as $\left[\bar{p}_{TTI}^{DP,91\%}, \bar{p}_{TTI}^{DP,100\%}\right]$, $\left[\bar{p}_{TTI}^{D,91\%}, \bar{p}_{TTI}^{D,100\%}\right]$ and $\left[\bar{p}_{TTI}^{P,91\%}, \bar{p}_{TTI}^{P,100\%}\right]$ are listed for the analyzed scheduling rules and policies when the windowing factor takes the considered discrete values. Also, the percentages of TTIs (mean and STD) of testing reward type for each objective (e.g., HoL delay, PDR and DP) are shown in detail for the same algorithms and windowing factor parameterizations. Based on the simulation results provided in Appendix H, the optimum windowing factor for both CBR and VBR traffic types belongs to the domain of $\rho \in [5.5, 300]$. For these values, the mean percentages of DP feasible TTIs are maximized and the percentages of punishment rewards are minimized.

7.3 DSR-CMOO MDP Focusing on HoL Delay, PDR, GBR and NGMN Fairness Objectives

The windowing factor affects the scheduling performances for both GBR and PDR objectives especially when large values $(\rho > 50)$ are involved in the optimization problems. As seen, when large windowing factors are used in the DSR-SMOO MDP problems focusing on the GBR objective, the mean percentage of feasible TTIs $\overline{p}_{TTI}^{G,100\%}$ increases but the reward function does not sense the immediate effect of the applied action in a given state. The mean percentage of feasible TTIs $\overline{p}_{TTI}^{P,100\%}$ from the viewpoint of the PDR objective can be improved when the windowing factor decreases. Moreover, from the NGMN requirement point of view, the number of feasible TTIs can be improved under different RL approaches when the windowing factor belongs to the optimum interval of $\rho \in [3.0; 4.0]$ and when the considered traffic is the full buffer model.

When the DSR-CMOO problem needs to be solved at each TTI in terms of the NGMN requirement, GBR, HoL delay and PDR objectives, the optimum range of windowing factor should be found at each TTI in order to maximize the multi-objective reward. In this sense, in this section is introduced an improved version of CACLA2, which is able to adapt three parameters $(\alpha_t, \beta_t, \rho_t)$ in order to improve the convergence of the DSR-CMOO MDP problems to the terminal states. The newest approach is entitled CACLA2+ which uses the fairness agent shown in Fig. 5.10 from Chapter 5. The fairness MLPNN function is trained based on the fairness observations which are extracted from the entire controller state space. The rest of this section is organized as follows: Sub-section 7.3.1 presents the DSR-CMOO optimization model, Sub-section 7.3.2 presents the novel RL approach which adapts the windowing factor each time when the fairness agent is selected by the QoS agent, Sub-section 7.3.3 presents the controller state space elements, Sub-section 7.3.4 proposes the multi-objective reward function and finally, Sub-section 7.3.5 analyzes the performance of the obtained sustainable scheduling policies for both CBR and VBR traffic types.

7.3.1 The Optimization Problem

The DSR-CMOO MDP problems focusing on the NGMN fairness requirement, GBR, HoL packet delay and the PDR objectives consider the scheduling rules introduced in Chapters 3 and 6 as follows:

1. NGMN requirement objective: GPF-DP;
2. GBR objective: GPF-BF, GPF-RAD, GPF-mM and GPF-LM;
3. HoL delay objective: GPF-EDF, GPF-MLWDF, GPD-LOG, GPF EXP1 and GPF-EXP2;
4. PDR objective: GPF-PLF and GPF-OPLF;
5. Queue stability objective: modified version of GPF-MDU;

Before going through more precise details, some acronyms used for the multi-objective combinations are specified below:

- **FGDP**: NGMN **F**airness Requirement, **G**BR, HoL **D**elay and **P**DR Objectives;
- **FG**: NGMN **F**airness Requirement and **G**BR objectives;
- **DP**: HoL packet **D**elay and **P**DR objectives;
- **GDP**: **G**BR, HoL **D**elay and **P**DR objectives;
- **GD**: **G**BR and HoL **D**elay objectives;

When combining the presented scheduling rules, the obtained DSR-CMOO $\left(P_{DP}^{FG}\right)$ problem is highlighted by Eq. 7.16.a and Eq. 7.16.b where $\left\{c_{2,1}[t],c_{3,g}[t],c_{4,d}[t],c_{5,p}[t],c_{6,1}[t]\right\}$, $\forall(g,d,p)=1,..,(4,5,2)$ represents the QoS controller action index which is mapped in the scheduling decision for the optimization problem. The set of vectors $\left\{u_{1,i}^{2}[t],u_{g,i}^{3}[t],u_{p,i}^{4}[t],u_{d,i}^{5}[t],u_{1,i}^{6}[t]\right\}$ represents the MU assignment variable which allocates the same MU function to each user $i\in\mathcal{U}_t$ at each TTI t.

By following the Equations 7.16.a, 7.16.b, and 7.16.c, the QoS controller should take optimal actions $\mathcal{A}_t^{a,FGDP}$, $\forall a=1,..,13$ at each TTI in order to maximize the concurrent optimization problem $\left(P_{DP}^{FG}\right)$, to respect the set of

$$\left(P_{DP}^{FG}\right):\max_{\pi_{RB}[t]}\Bigg\{c_{2,1}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{1,i}^{2}[t]\cdot b_{i,j}[t]\cdot\frac{\left(r_{i,j}[t]\right)^{\beta_t}}{\left(\overline{T}_i[t]\right)^{\alpha_t}}\right]\ 1:(GPF-DP)$$

$$+c_{3,1}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{1,i}^{3}[t]\cdot b_{i,j}[t]\cdot\left(1+\omega_{1,1}^{3}\cdot e^{-\omega_{1,2}^{3}\cdot\left(\overline{\overline{T}}_i[t]-\breve{T}_i[t]\right)}\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 2:\begin{pmatrix}GPF-\\BF\end{pmatrix}$$

$$+c_{3,2}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{2,i}^{3}[t]\cdot b_{i,j}[t]\cdot e^{\omega_{3,i}^{3}\cdot TC_i[t]}\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 3:(GPF-mM)$$

$$+c_{3,3}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{3,i}^{3}[t]\cdot b_{i,j}[t]\cdot\frac{\breve{T}_i[t]}{\overline{\overline{T}}_i[t]}\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 4:(GPF-RAD)$$

$$+c_{3,4}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{4,i}^{3}[t]\cdot b_{i,j}[t]\cdot log\left(\omega_{4,i}^{3}+\lambda_i\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\Bigg\}\ 5:(GPF-LM)$$

$$+c_{4,1}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{1,i}^{4}[t]\cdot b_{i,j}[t]\cdot\omega_{1,i}^{4}\cdot d_i^{HoL}[t]\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 6:(GPF-MLWDF)$$

$$+c_{4,2}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{2,i}^{4}[t]\cdot b_{i,j}[t]\cdot exp\left(\frac{\omega_{2,i}^{4}\cdot d_i^{HoL}[t]}{1+\sqrt{d_{ik}^{HoL}[t]}}\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 7:\begin{pmatrix}GPF-\\EXP1\end{pmatrix}$$

$$+c_{4,3}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{3,i}^{4}[t]\cdot b_{i,j}[t]\cdot exp\left(\frac{\omega_{3,i}^{4}\cdot d_i^{HoL}[t]-\widehat{d^{HoL}}[t]}{1+\sqrt{\widehat{d^{HoL}}[t]}}\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 8:\begin{pmatrix}GPF-\\EXP2\end{pmatrix}$$

$$+c_{4,4}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{4,i}^{4}[t]\cdot b_{i,j}[t]\cdot log\left(\omega_{4,1}^{4}+\omega_{4,2}^{4}\cdot\frac{d_i^{HoL}[t]}{\breve{d}_{i,L}^{HoL}[t]}\right)\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 9:\begin{pmatrix}GPF-\\LOG\end{pmatrix}$$

$$+c_{4,5}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{5,i}^{4}[t]\cdot b_{i,j}[t]\cdot\frac{1}{\breve{d}_{i,L}^{HoL}[t]-d_i^{HoL}[t]}\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 10:\begin{pmatrix}GPF-\\EDF\end{pmatrix}$$

$$+c_{5,1}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{1,i}^{5}[t]\cdot b_{i,j}[t]\cdot R_i^{PL}[t]\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 11:(GPF-PLF)$$

$$+c_{5,2}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{2,i}^{5}[t]\cdot b_{i,j}[t]\cdot\frac{R_i^{PL}[t]\cdot d_i^{HoL}[t]}{\breve{R}_i^{PL}[t]\cdot\breve{d}_{i,L}^{HoL}[t]}\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\ 12:(GPF-OPLF)$$

$$+c_{6,1}[t]\cdot\left[\sum_{i=1}^{|\mathcal{U}_t|}\sum_{j=1}^{|\mathcal{B}|}u_{1,i}^{6}[t]\cdot b_{i,j}[t]\cdot\overline{q_i^{TX}}[t]\cdot\frac{\left(r_{i,j}[t]\right)^{\beta}}{\left(\overline{T}_i[t]\right)^{\alpha}}\right]\Bigg\}\ 13:(GPF-MDU)$$

(7.16.a)

$$
\begin{aligned}
& c_{2,1}[t]+\sum_{w_3=1}^{4} c_{3,w_3}[t]+\sum_{w_4=1}^{5} c_{4,w_4}[t]+\sum_{w_5=1}^{2} c_{5,w_5}[t]+c_{6,1}[t]=1 \\
& u_{1,i}^{2}[t]+\sum_{w_3=1}^{4} u_{w_3,i}^{3}[t]+\sum_{w_4=1}^{5} u_{w_4,i}^{4}[t]+\sum_{w_5=1}^{2} u_{w_5,i}^{5}[t]+u_{1,i}^{6}[t]=1, i=1,...,|\mathcal{U}_t| \\
& \sum_{i=1}^{|\mathcal{U}_t|} u_{w_o^*,i}^{o}[t]=|\mathcal{U}_t|, w_o^* \in \mathcal{PU}_o, o=\{2,3,4,5,6\} \\
\left(C_{DP}^{FG}\right)\ s.t.: & \sum_{i=1}^{|\mathcal{U}_t|} u_{w_o^{\otimes},i}^{o}[t]=0, w_o^{\otimes}=1,...,|\mathcal{PU}_o|, o=\{2,3,4,5,6\}, \forall w_o^{\otimes} \neq w_o^* \\
& \sum_{i=1}^{|\mathcal{U}_t|} b_{i,j}[t]=1,\ \ j=1,...,|\mathcal{B}| \\
& \left\{c_{2,1}[t], c_{3,w_3}[t], c_{4,w_4}[t], c_{5,w_5}[t], c_{6,1}[t]\right\} \in \{0,1\}, \forall \{w_3, w_4, w_5\} \\
& \left\{u_{1,i}^{2}[t], u_{w_3,i}^{3}[t], u_{w_4,i}^{4}[t], u_{w_5,i}^{5}[t], u_{1,i}^{6}[t]\right\} \in \{0,1\}, \forall \{w_3, w_4, w_5\}, \forall i \in \mathcal{U}_t \\
& b_{i,j}[t] \in \{0,1\}, \forall i \in \mathcal{U}_t, \forall j \in \mathcal{B}
\end{aligned}
\tag{7.16.b}
$$

$$
\begin{aligned}
& \psi\left(\overline{\overline{T}}_i[t]\right) \leq \overline{\overline{T}}_i[t], \forall i \in \mathcal{U}_t \\
& \overline{\overline{T}}_i[t] \geq \underline{\underline{T}}_i[t], \forall i \in \mathcal{U}_t \\
\left(O_{DP}^{FG}\right): & d_i^{HoL}[t] \leq \overline{d}_{i,L}^{HoL}[t], \forall i \in \mathcal{U}_t \\
& R_i^{PL}[t] \leq \overline{R}_i^{PL}[t], \forall i \in \mathcal{U}_t
\end{aligned}
\tag{7.16.c}
$$

convex constraints $\left(C_{DP}^{FG}\right)$ and to respect the set of objective constraints $\left(O_{DP}^{FG}\right)$ in order to maximize the total scheduler return. If $\mathcal{A}_t^{1,FGDP}=1$ is selected, the GPF-DP scheduling rule is applied and implicitly, the NGMN fairness objective is addressed. When all objective constraints are satisfied, all active bearers are 100% satisfied from the viewpoints of the NGMN fairness, GBR, HoL delay and PDR objectives. The analyzed CMOO problem from Eq. 7.16 is sub-optimal and it is considered to be a linear optimization programming model since the scheduling vector $c_{o,w_o}[t]$ is already known based on the controller output decision for a given controller state $\mathcal{S}_t^C$. Moreover, the fairness parameters for the entire set of scheduling rules excepting GPF-DP is fixed to $(\alpha=1,\beta=1)$. If the GPF-DP rule is selected, the parameters (α_t,β_t) are adapted based on the CACLA2 RL.

When the aforementioned FGDP problems are optimized TTI-by-TTI, the performance of other combined objectives is studied in order to highlight the advantages of using the proposed scheduling policies instead of the existing scheduling rules. From the architectural point of view, the GPF-DP parameterization is performed by using the fairness controller as shown in Fig. 5.10 in Chapter 5. Alongside of fairness parameter adaptations, the novel fairness controller is able to parameterize the windowing factor by using the CACLA2+ RL approach. The adaptation period of the windowing factor depends exclusively on the QoS agent when decides to select the fairness agent.

7.3.2 Continuous Actor Critic Learning Automata with a Dynamic Windowing Factor

The QoS controller selects the scheduling rule at TTI *t* based on the FGDP DSR-CMOO MDP problem $\left(\mathcal{P}_{FGDP}^{MDP}\right)$ defined by the following set:

$$\left(\mathcal{P}_{FGDP}^{MDP}\right):\left[\mathcal{S}_{t-1}^{C,FGDP},\mathcal{A}_{t-1}^{a,FGDP},\mathcal{RW}_{t}^{FGDP},\mathcal{S}_{t}^{C,FGDP},\mathcal{A}_{t}^{a,FGDP}\right]_{a=1,\ldots,|\mathcal{A}|} \tag{7.17}$$

where $\mathcal{RW}_{t}^{FGDP}$ represents the scheduler reward based on the combined FGDP objectives. When the action $\mathcal{A}_{t}^{1,FGDP}$ is selected during the exploration/exploitation period, the corresponding scheduling rule is the GPF scheme with the double parameterization. This way, CACLA2 RL provides the necessary parameter steps $\left(\Delta\alpha_t,\Delta\beta_t\right)$ in order to satisfy the NGMN fairness requirement. When the optimum windowing factor must be reached from the online exploration/ exploitation periods, CACLA2+ adapts three parameters $\left(\Delta\alpha_t,\Delta\beta_t,\Delta\rho_t\right)$, where $\Delta\rho_t$ represents the necessary windowing factor step at TTI *t*. Therefore, the action set for CACLA2+ RL algorithm is represented by Eq. 7.18:

$$\left(\mathcal{A}_{t}^{F}\right):\begin{cases}\alpha_t=\alpha_{t-p}+\Delta\alpha_t\\ \beta_t=\beta_{t-p}+\Delta\beta_t\\ \rho_t=\rho_{t-p}+\Delta\rho_t\end{cases} \tag{7.18}$$

where p is the time instant when the fairness MLPNN weights were updated last time. When the parameter ρ_t is updated, the newest time window domain impacts the NGMN fairness, GBR and PDR objectives until the GPF-DP scheduling rule is selected again. The DSR-SMOO MDP problem focusing on the NGMN requirement for the CACLA2+ RL algorithm is defined by using Eq. 7.19:

$$\left(\mathcal{P}_F^{MDP}\right):\left[\mathcal{S}_{t-p}^{C,F},\mathcal{A}_{t-p}^{F},\mathcal{RW}_t^{FDP},\mathcal{S}_t^{C,F},\mathcal{A}_t^{F}\right]_{\mathcal{A}_t^F\in\mathbb{R}^3_{[-1,1]}} \tag{7.19}$$

where $\mathcal{S}_t^{C,F}\subset\mathcal{S}_t^{C,FGDP}$ is the fairness controller state space which is a subset from the overall controller state space being focused on the FGDP multi-objective criterion. The considered subspace for CACLA2+ follows the form of Eq. 6.7.b with a slight modification as indicated by Eq. 7.20:

$$\mathcal{S}_t^{C,FDP}=\left\{\alpha_{t-p},\beta_{t-p},\rho_{t-p},N_{CL}^{A},\sigma_{CL}^{t},ks_t,d_{ks}^{t},\mu_T^t,\sigma_T^t,N_U^t\right\} \tag{7.20}$$

The NGMN fairness reward $\mathcal{RW}_t^{FDP}$ remains unchanged and it is considered by the entire DSR-CMOO problem at each TTI and when the fairness MLPNN weights are updated. Even if the dynamic parameterization of the windowing factor affects the rewards of other objectives at each TTI t, CACLA2+ RL algorithm trains the MLPNN weights based only on $\mathcal{RW}_t^{FDP}$ which is received at TTI t as a consequence of applying the fairness action $\mathcal{A}_{t-p}^{F}$ at TTI *t-p* in the state $\mathcal{S}_{t-p}^{C,F}$. In other words, CACLA2+ trains the MLPNN functions at different time scales when compared with other RL approaches which train the QoS MLPNN functions TTI-by-TTI. For these reasons, in the exploration stage, the period of policy improvement should be much larger than the period of policy evaluation in order to increase the number of updates when the fairness MLPNN functions are trained and to provide more chances for the GPF-DP discipline to be selected.

7.3.3 Controller State Space

The controller state space considers the entire set of observations for each discussed objective in order to localize the feasible states when the total scheduler reward is maximized. Let us define $\mathcal{S}_t^{C,FGDP}\in\mathcal{FAFGDP}$ the feasible set of state

observations when all the objectives are satisfied and $\mathcal{S}_t^{C,FGDP} \in \mathcal{UFFGDP}$ the unfeasible region where at least one objective is not satisfied. Then, the FGDP state feasibility can be defined according to Eq. 7.21:

$$\mathcal{S}_t^{C,FGDP} \in \begin{cases} \{\mathcal{FAFGDP}\},\ if\ d_i^{HOL}[t] \le d_{i,L}^{\overline{HoL}}[t]\ and\ R_i^{PL}[t] \le R_i^{\overline{PL}}[t] \\ \qquad and\ \overline{\overline{T}}_i[t] \ge \underline{T}_i[t]\ and\ \psi\left(\overline{\overline{T}}_i[t]\right) \le \overline{\overline{T}}_i[t], \forall i \in \mathcal{U}_t \\ \{\mathcal{UFFGDP}\},\ if\ \exists d_i^{HOL}[t] > d_{i,L}^{\overline{HoL}}[t]\ or\ \exists R_i^{PL}[t] > R_i^{\overline{PL}}[t] \\ \qquad or\ \exists \overline{\overline{T}}_i[t] < \underline{T}_i[t]\ or\ \exists \psi\left(\overline{\overline{T}}_i[t]\right) > \overline{\overline{T}}_i[t], \forall i \in \mathcal{U}_t \end{cases} \tag{7.21}$$

By unifying the state spaces for the NGMN fairness requirement, GBR, HoL delay and PDR objectives, the DSR-CMOO MDP problems use the following data set in order to train the QoS structure of MLPNN functions at each TTI *t*:

$$\begin{aligned} \mathcal{S}_t^{C,FGDP} = \{ & \mathcal{A}_{t-1}^{a,FGDP}, \alpha_{t-p}, \beta_{t-p}, \rho_{t-p}, N_{CL}^{A}, \sigma_{CL}^{t}, ks_t, d_{ks}^{t}, \mu_{\overline{T}}^{t}, \sigma_{\overline{T}}^{t}, |\mathcal{U}_t|, \quad (1) \\ & \mu_{\overline{\overline{T}}}^{t}, \sigma_{\overline{\overline{T}}}^{t}, \mu_{\lambda_{\overline{\overline{T}}}^{G}}^{t}, \sigma_{\lambda_{\overline{\overline{T}}}^{G}}^{t}, N_{t,G}^{SAT}, N_{t,G}^{USAT}, \underline{T}_t, \qquad (2) \\ & \mu_{D}^{t}, \sigma_{D}^{t}, \mu_{\lambda D}^{t}, \sigma_{\lambda D}^{t}, N_{t,D}^{SAT}, N_{t,D}^{USAT}, d_t^{\overline{HoL}}, \quad (3) \\ & \mu_{P}^{t}, \sigma_{P}^{t}, \mu_{\lambda P}^{t}, \sigma_{\lambda P}^{t}, N_{t,P}^{SAT}, N_{t,P}^{USAT}, R_t^{\overline{PL}}, \qquad (4) \\ & \mu_{qTX}^{t}, \sigma_{qTX}^{t}, \mu_{\overline{\lambda}}^{t}, \sigma_{\overline{\lambda}}^{t}, N_{t,Queues}^{Active}, N_{t,Queues}^{Unactive} \} \qquad (5) \end{aligned} \tag{7.22}$$

where all the observations were introduced gradually in the current study. Basically, when the normalized numbers of unsatisfied bearers from the viewpoints of GBR, HoL delay and PDR objectives are $\{N_{t,G}^{USAT}, N_{t,D}^{USAT}, N_{t,P}^{USAT}\} = 0$ and the minimum distance in the CDF domain is less than the fairness confidence parameter such as $d_{ks}^{t} < \xi$, then the controller state is considered to be feasible at TTI *t* and consequently, the state belongs to $\mathcal{S}_t^{C,FGDP} \in \mathcal{FAFGDP}$. Otherwise, if one of these conditions is not respected, then the controller state is considered unfeasible and $\mathcal{S}_t^{C,FGDP} \in \mathcal{UFFGDP}$. The state space defined by Eq. 7.22 requires higher number of MLPNN nodes for the hidden layer in order to enhance the approximations between each state and the output MLPNN decision which selects the scheduling rule for the DSR-CMOO combinatorial problem.

7.3.4 Reward Function

The multi-objective reward function $\mathcal{RW}_t^{FGDP}$ considers the weighted sum of each intrinsic objective reward as described in Eq. 7.23:

$$\begin{aligned}\mathcal{RWI}_t^{FGDP}\left(\mathcal{A}_{t-1}^{a,FGDP},\mathcal{S}_{t-1}^{C,FGDP}\right) = \delta_F \cdot \mathcal{RW}_t^{F}\left(\mathcal{A}_{t-1}^{a,FGDP},\mathcal{S}_{t-1}^{C,FGDP}\right) \\ +\delta_G \cdot \mathcal{RW}_t^{G}\left(\mathcal{A}_{t-1}^{a,FGDP},\mathcal{S}_{t-1}^{C,FGDP}\right) \\ +\delta_D \cdot \mathcal{RW}_t^{D}\left(\mathcal{A}_{t-1}^{a,FGDP},\mathcal{S}_{t-1}^{C,FGDP}\right) \\ +\delta_P \cdot \mathcal{RW}_t^{P}\left(\mathcal{A}_{t-1}^{a,FGDP},\mathcal{S}_{t-1}^{C,FGDP}\right)\end{aligned} \tag{7.23}$$

where the tuple $\left(\delta_F,\delta_G,\delta_D,\delta_P\right)\in[0,1]$ represents the importance of each objective in the considered CMOO problem by imposing the constraint of $\delta_F+\delta_G+\delta_D+\delta_P=1$. The weights can be selected based on the used traffic class. For instance, for the VoIP services, the DSR-CMOO combinatorial problem can use $\left(\delta_F=0,\delta_G=0.5,\delta_D=0.5,\delta_P=0\right)$ since the most important objective is to provide the requested services in some given HoL delay and GBR constraints.

Another important set of parameters is $\left(\ell_G,\ell_D,\ell_P\right)$ which permits to select the objective rewards as a temporal difference between two consecutive intrinsic rewards. For the FGDP grand objective, an additional parameter such as ℓ_{FGDP} is needed in order to set the global reward function as defined by Eq. 7.24:

$$\mathcal{RW}^{FGDP}[t]=\begin{cases}1, & if\ \mathcal{RWI}^{FGDP}[t]=1\\ \mathcal{RWI}^{FGDP}[t]-\ell_{FGDP}\cdot\mathcal{RWI}^{FGDP}[t-1], & otherwise\end{cases} \tag{7.24}$$

Based on Eq. 7.24, when $\ell_{FGDP}=0$, only the current intrinsic reward is used in order to compute the overall FGDP reward. When $\ell_{FGDP}=1$, the FGDP reward considers the improvement between two consecutive intrinsic rewards. From the perspective of the DSR-CMOO MDP problems, both types of parameter sets, $\left(\delta_F,\delta_G,\delta_D,\delta_P\right)$ and $\left(\ell_G,\ell_D,\ell_P,\ell_{FGDP}\right)$, are crucial in finding the optimal scheduling rules at each TTI. For the fairness objective, by definition, the reward function is computed based on the difference of fairness parameters between consecutive controller states as shown in Equations 6.17, 6.18, 6.19.a and 6.19.b.

7.3.5 Performance Evaluation of Sustainable Scheduling Policies Focusing on NGMN Fairness, GBR, HoL Delay and PDR Objectives

The performance evaluation of FGDP scheduling policies are conducted through variable windowing factors which are decided by the CACLA2+ agent for the CBR and VBR traffic types. The rest of this sub-section is organized as follows: Sub-section 7.3.5.1 introduces the simulation scenario, Sub-section 7.3.5.2 presents the results of the sustainable policies for the CBR traffic type and finally, Sub-section 7.3.5.3 evaluates the learned policies for the VBR traffic type.

7.3.5.1 Simulation Scenario

Both traffic types of CBR and VBR are analyzed for the CMOO MDP problems focusing on the NGMN fairness, GBR, HoL delay and PDR objectives. The parameter settings for the LTE scheduler respect the configuration provided in Table 7.1. During the exploration/exploitation stages, the traffic load and the QoS requirements are changed randomly at each 1000 TTIs. The multi-objective reward $\mathcal{RW}_t^{FGDP}$ considers the difference between consecutive intrinsic rewards $\left(\ell_{FGDP}=1\right)$ whereas other QoS rewards take into account only the instantaneous intrinsic reward $\left(\ell_G=0,\ell_D=0,\ell_P=0\right)$. For the NGMN fairness evaluation, the confidence factor is increased to the value of $\xi=0.35$ in order to enhance the number of episodes during the exploration stage. Equal weights are considered for the QoS objectives when the global reward is computed $\left(\delta_F=0.25,\delta_G=0.25,\right.$ $\left.\delta_D=0.25,\delta_P=0.25\right)$. This way, all objectives have the same importance when training the scheduling policies. When the HoL delay objective performance is evaluated, the percentage of satisfied users is matched against the lower delay constraint with $\vartheta_D=0.1$. The NGMN fairness, GBR and PDR objectives are evaluated based on a dynamic windowing factor which is selected in the interval of [2.5; 50] by the continuous action of CACLA2+. The rest of the parameters for the QoS and fairness MLPNN functions are highlighted in Table 7.1. Table 7.2

Table 7.1. LTE Scheduler Controller Parameters for DSR-CMOO Focusing on FGDP Objectives

Parameters Name	Description/Values
System Bandwidth/Cell Radius	20 MHz/1000m [36]
User Speed/Mobility Model	120Kmph/Random Direction
Channel Model	Jakes Model (Appendix B)
Path Loss / Penetration Loss	Macro Cell Model / 10 dB [36]
Interfered Cells/ Shadowing STD	0/8dB [36]
Carrier Frequency/DL Power	2GHz/43dBm [36]
Frame Structure	FDD
CQI Reporting Mode	Full-band, periodic at each TTI
PUCCH Model	Errorless
Scheduler Type	GPF-DP[85], GPF-BF[99], GPF-RAD [94], GPF-mM[82], GPF-LM(Novelty), GPF-EDF[113], GPF-MLWDF[55], GPF-LOG[110], GPF-EXP1[108], GPF-EXP2[110], GPF-PLF[56], GPF-OPLF[56], GPF-MDU[49]
$\left\{\omega_{1,1}^{3};\omega_{1,2}^{3};\omega_{2,i}^{3};\omega_{4,i}^{3}\right\}$	$\left\{1.25[99];13.1\cdot10^{-5}[99];10.1[82];2\right\}$
$\left\{\omega_{4,1}^{4},\omega_{4,2}^{4},\omega_{2,i}^{4}\right\}$	$\left\{1.1[110],5.0[110],6.0[108]\right\}$
$\left\{\omega_{1,i}^{4},\omega_{3,i}^{4}\right\}$	$-log_{10}\,\bar{R}_{i}^{PL}/\bar{d}_{i,L}^{HoL}$ [55], [110]
Traffic Type	CBR/VBR
Max. Number of schedulable users $\left(N_{Sched}^{Max}\right)$ at each TTI	10 (Optimum)
RLC ARQ	Acknowledged Mode (Maximum 5 retransmissions)
AMC Levels	QPSK (1/3, 1/2, 2/3), 16-QAM (1/2, 2/3, 5/6) 64-QAM (2/3, 5/6) [36]
Target BLER	10% (Appendix B)
Number of Users $\left(\lvert\mathcal{U}\rvert\right)$	Variable(Exploration) : 15-120 Variable(Exploitation) : 15-80
RL Algorithms For Discrete Actions (QoS Objectives)	Q-L, DoubleQ-L, SARSA, QV, QV2, QVMAX, QVMAX2, ACLA
Discrete QoS MLPNN Actions	1:GPF-DP, 2:GPF-BF, 3:GPF-RAD, 4:GPF-mM, 5:GPF-LM,6:GPF-EDF, 7:GPF-MLWDF, 8:GPF-LOG, 9 :GPF-EXP1, 10:GPF-EXP2, 11:GPF-PLF, 12:GPF-OPLF, 13:GPF-MDU
Controller Timescale	1 TTI
$\left\{\delta_{F},\delta_{G},\delta_{D},\delta_{P}\right\}$	$\left\{0.25,0.25,0.25,0.25\right\}$
$\left\{\ell_{G},\ell_{D},\ell_{P},\ell_{FGDP}\right\}$	$\left\{0,0,0,1\right\}$
Number of MLPNN layers / Activation Functions	3/ input layer: linear activation,

(fairness and QoS objectives)	hidden layer: tangent hyperbolic activation, output layer: linear activation
Number of Hidden Nodes for QoS MLPNN	150 (Optimum)
RL for GFP-DP and Dynamic Windowing Factor	CACLA2+
Continuous NGMN MLPNN Actions	$\mathcal{A}_t^{FDP} = (\Delta\alpha_t, \Delta\beta_t, \Delta\rho_t) \in \mathbb{R}^3_{[-1,1]}$
Number of Hidden nodes For Fairness MLPNN	50 (Optimum)
Exploration/Exploitation Periods	500/100 (optimum)
Windowing Factor (AUT-MMF and PDR)	Continuous based on CACLA2+ decision: $\rho \in [2.5; 50]$
Dynamic GBR Constraints	$\overline{T} = \{32; 64; 128; 256; 512; 1024\}\, kbps$
Dynamic HoL Delay Constraints	$d_i^{\overline{HoL}} \in \{50, 100, 150, 200, 250, 300\}\, ms,\ \vartheta_D = 0.1$
Dynamic PDR Constraints	$R_i^{\overline{PL}} \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
CQI Aggregation Schemes	Top CQI Mass Modes $\{Top3\} : N_{CT} = \{64\}$
CBR Traffic Type	Data Rates based on GBR Constraints $\lambda = \{32; 64; 128; 256; 512; 1024\}\, kbps$
VBR Traffic Type	Packet size: Pareto Distrib. $(x = 35.5; \alpha = 1.1)$ [13] Arriv. Rate: Geometric Distrib. $(\mu = 1.5; \sigma = 1.93)$ [13]

shows the set of parameters which is used by the proposed RL algorithms. As shown in Table 7.2, by increasing the Boltzmann parameters and decreasing the greedy probability threshold, the GPF-DP scheduling rule has higher chances to be selected in the exploration period, and consequently, CACLA2+ RL algorithm is able to reinforce the fairness MLPNN structure for enough epochs in order to be able to provide the optimal continuous actions during the exploitation period.

In the exploitation period, the policies are tested by using 10 simulations with different scenarios in order to compute the mean and STD parameters for the obtained results. Then, the scheduling performance is evaluated in terms of the mean percentage of TTIs for different satisfaction levels when the QoS objectives are used in different combinations such that: $\overline{p}_{TTI}^{FG,x\%}$, $\overline{p}_{TTI}^{DP,x\%}$, $\overline{p}_{TTI}^{GD,x\%}$, $\overline{p}_{TTI}^{GDP,x\%}$ and $\overline{p}_{TTI}^{FGDP,x\%}$. It is important to note that the percentage $\overline{p}_{TTI}^{FGDP,x\%}$ is updated at each TTI if and only if the scheduler respects the NGMN requirement. The main focus

Table 7.2 RL Parameters for DSR-CMOO Focusing on FGDP Objectives

RL Algorithm (Fairness SMOO)	Learning Rates for Action Values (η^Q)	Learning Rates for State Values (η^V)	Discount Factor (γ)	Exploration Type (ε,τ)
Q	0.0001	-	0.99	Greedy $(\varepsilon = 5\cdot 10^{-4})$
DoubleQ	0.0001	-	0.99	Greedy $(\varepsilon = 5\cdot 10^{-4})$
SARSA	0.001	-	0.99	Boltzmann $(\tau = 10)$
QV	0.01	0.0001	0.99	Boltzmann $(\tau = 10)$
QV2	0.01	0.0001	0.95	Boltzmann $(\tau = 10)$
QVMAX	0.01	0.0001	0.99	Boltzmann $(\tau = 10)$
QVMAX2	0.01	0.0001	0.95	Boltzmann $(\tau = 10)$
ACLA	0.0001	0.0001	0.99	Greedy $(\varepsilon = 5\cdot 10^{-4})$
CACLA2+	0.01	0.01	0.99	Greedy $(\varepsilon = 0.5)$

is to maximize the percentage of TTIs when all active bearers are satisfied from the viewpoint of the combined FGDP objectives. The obtained scheduling policies are evaluated in terms of the mean percentage of TTIs for different reward types under different objective combinations. The idea is to minimize at the same time the amount of punishment and moderate rewards in the exploitation stage for a better sustainability of the obtained set of scheduling policies.

7.3.5.2 DSR-CMOO Focusing on HoL Delay, PDR, GBR and NGMN Fairness Objectives with the CBR Traffic Type

Figure 7.17 analyzes the performances of the learned scheduling policies from the viewpoint of the mean percentage of GBR feasible TTIs, when compared with the classical scheduling rule, when the FGDP DSR-CMOO problem is performed. The SARSA algorithm provides the best performance in the domain of $\left[\bar{p}_{TTI}^{G,94\%}, \bar{p}_{TTI}^{G,100\%}\right]$ whereas the DQ-learning selects the best scheduling rules for the GBR objective in the performance interval of $\left[\bar{p}_{TTI}^{G,80\%}, \bar{p}_{TTI}^{G,94\%}\right]$. For the CBR traffic type, the GPF-LOG and GPF-MDU rules perform better than the GBR oriented scheduling rules such as GPF-BF and GPF-RAD. By focusing on certain target

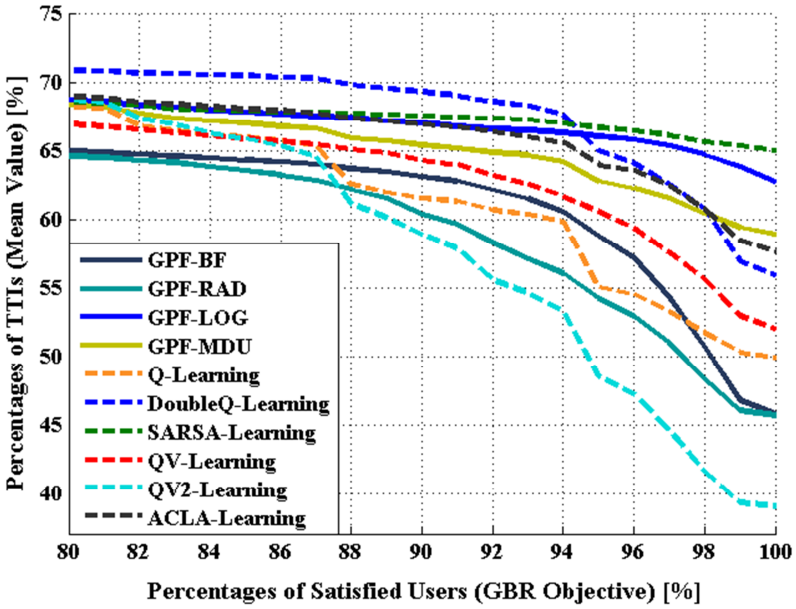


Fig. 7.17 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

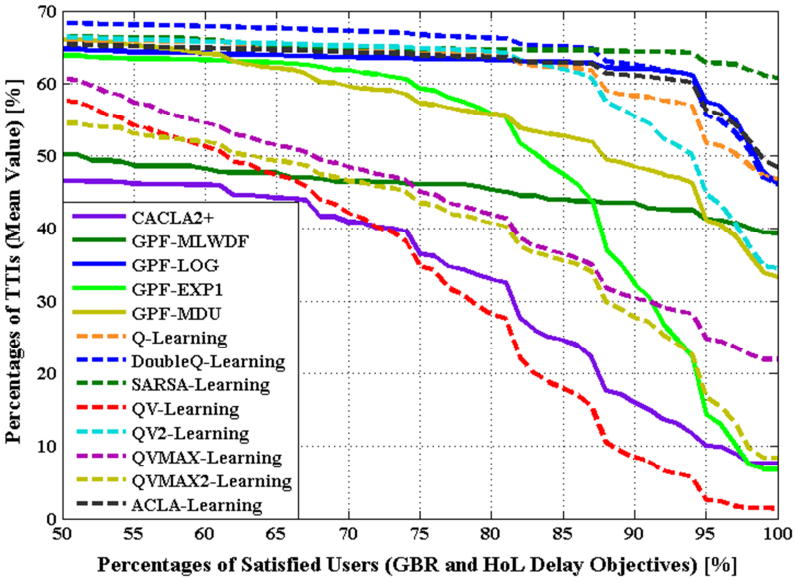


Fig. 7.18 Mean Percentages of TTIs vs. Percentages of GD Satisfied Bearers for the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

delay, the GBR constraints can be satisfied much better by the scheduling rules focusing on HoL delay when the CBR traffic type is considered. From the viewpoint of GD objectives (Fig. 7.18), SARSA is the best choice when the interval of $\left[\overline{p}_{TTI}^{GD,87\%}, \overline{p}_{TTI}^{GD,100\%}\right]$ is considered, and DoubleQ-learning performs the

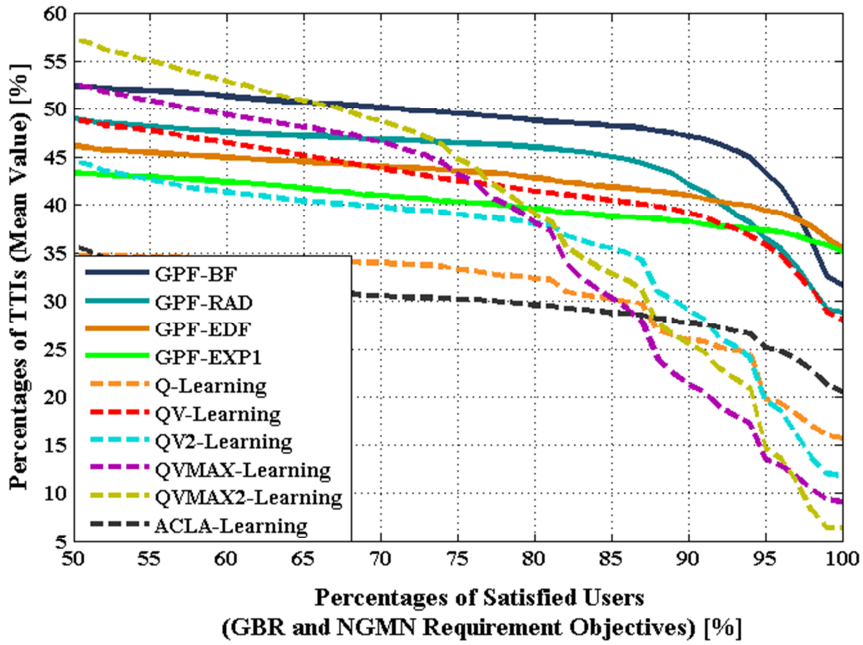


Fig. 7.19 Mean Percentages of TTIs vs. Percentages of FG Satisfied Bearers for the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

best for the rest of the considered interval. SARSA indicates more than 10% gain from the percentage $\bar{p}_{TTI}^{GD,100\%}$ point of view when compared with GPF-LOG, DoubleQ, ACLA or Q-learning algorithms. It is important to note that, the CACLA2+ scheduling policy which is oriented more on the fairness satisfaction, can achieve a plausible level of tradeoff between GBR and HoL delay objectives.

Figure 7.19 analyzes the performance of the mean percentage of TTIs for the NGMN fairness and GBR tradeoff satisfaction. By focusing on satisfying all objectives at each TTI *t*, the proposed RL algorithms are not able to increase the percentage of TTIs for the particular tradeoff of FG objectives. For this reason, the GPF-BF scheduling rule gains more than 8% from the $\bar{p}_{TTI}^{FG,85\%}$ point of view when compared with other scheduling policies. It is interesting to notice that GPF-EXP1 which provides unsatisfactory performance for the DP tradeoff is able to outperform the GBR oriented scheduling rules by about 3% when the performance of $\bar{p}_{TTI}^{FG,100\%}$ percentage is measured.

When the DP tradeoff is considered (Fig. 7.20), all RL policies excepting QVMAX and QVMAX2 learning outperform the classical scheduling rules such

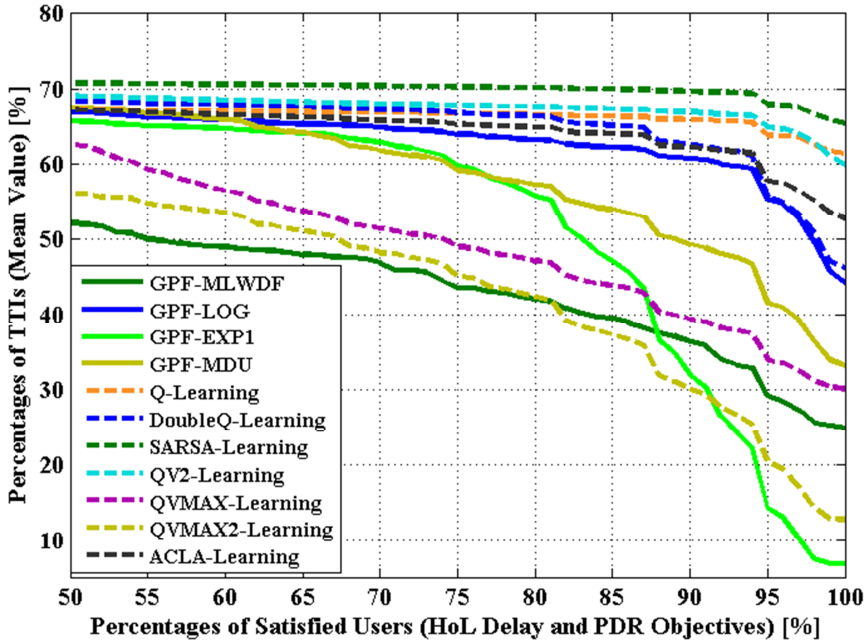


Fig. 7.20 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

as GPF-LOG, GPF-MDU, GPF-MLWDF or GPF-EXP1. The SARSA policy gains more than 20% in percentage of TTIs for $\overline{p}_{TTI}^{DP,100\%}$ when compared with GPF-LOG and more than 30% when matched against GPF-MDU. As expected, the worst performance from the viewpoint of DP objectives is shown by the GPF-EXP1 rule which obtains the highest STD factor in satisfying the radio bearers.

Figure 7.21 presents the results obtained in terms of the mean percentage of TTIs for satisfied bearers from the viewpoint of GDP objectives. The SARSA scheduling policy achieves more than 15% for the mean percentage of $\overline{p}_{TTI}^{GDP,100\%}$ when matched against the GPF-LOG rule and more than 25% when compared with the GPF-MDU scheduling technique. In fact, SARSA performs better than any other candidate in the performance domain of $\left[\overline{p}_{TTI}^{GDP,87\%}, \overline{p}_{TTI}^{GDP,100\%}\right]$. When the percentage of TTIs is analyzed in the domain of $\left[\overline{p}_{TTI}^{GDP,50\%}, \overline{p}_{TTI}^{GDP,87\%}\right]$, Q-L, DoubleQ, SARSA and ACLA provide a similar performance. QV, QVMAX and QVMAX2 policies show an unsatisfactory performance when compared with other scheduling rules such as GPF-EXP1, GPF-MDU or GPF-LOG.

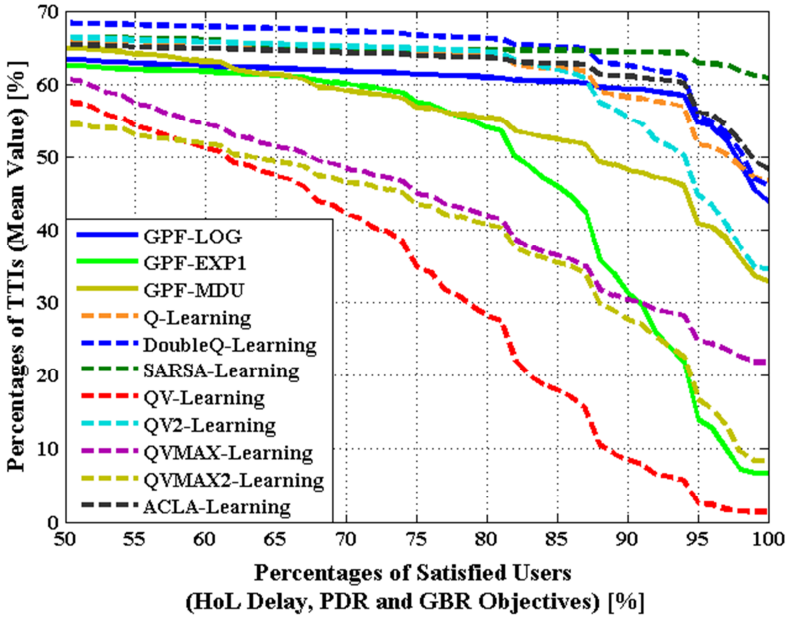


Fig. 7.21 Mean Percentages of TTIs vs. Percentages of GDP Satisfied Bearers for the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

The concurrent optimization problem exposed in Eq. 7.16 aims to maximize the mean percentage of TTIs from the perspective of a combined FGDP tradeoff satisfaction degree. The results provided in Fig. 7.22 indicate that one scheduling policy is not enough to increase the percentage of TTIs for the entire FGDP satisfaction domain. Therefore, SARSA policy performs the best when the performance domain of $\left[\overline{p}_{TTI}^{FGDP,98\%}, \overline{p}_{TTI}^{FGDP,100\%}\right]$ is considered. For the interval of $\left[\overline{p}_{TTI}^{FGDP,95\%}, \overline{p}_{TTI}^{FGDP,98\%}\right]$, the best choice is represented by the scheduling policy obtained when the ACLA actor-critic is performed. The DoubleQ policy performs the best when the satisfaction level of the active bearers belongs to the domain of $\left[\overline{p}_{TTI}^{FGDP,90\%}, \overline{p}_{TTI}^{FGDP,95\%}\right]$, and for a wider domain of $\left[\overline{p}_{TTI}^{FGDP,70\%}, \overline{p}_{TTI}^{FGDP,90\%}\right]$, the QV2 policy becomes the best choice when the tradeoff between all objectives is considered. The limitation of the GPF-LOG rule is denoted by the fact that a constant level in the FGDP multi-objective satisfaction is obtained for the interval of $\left[\overline{p}_{TTI}^{FGDP,50\%}, \overline{p}_{TTI}^{FGDP,95\%}\right]$ whereas GPF-EXP1 or GPF-MDU assures a much better performance when the considered domain of interest is $\left[\overline{p}_{TTI}^{FGDP,50\%}, \overline{p}_{TTI}^{FGDP,85\%}\right]$.

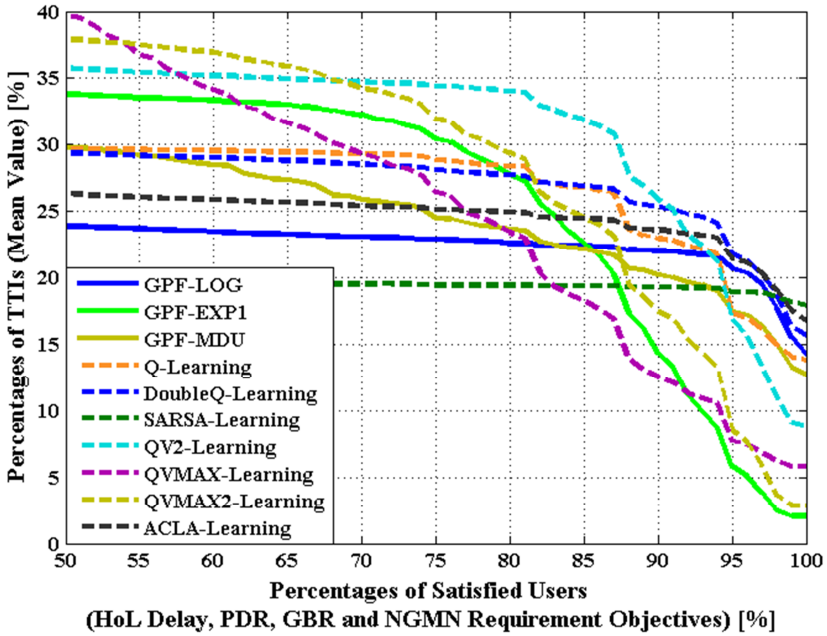


Fig. 7.22 Mean Percentages of TTIs vs. Percentages of FGDP Satisfied Bearers for the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

The mean percentage of TTIs for different reward types are measured for each QoS objective reward and for the global reward function proposed in Eq. 7.24. It is important to remind that only the global reward considers the temporal difference of the consecutive intrinsic rewards whereas other particular QoS rewards use the instantaneous QoS intrinsic rewards obtained at each TTI. From the GBR objective point of view (Fig. 7.23), the highest amount of maximum rewards is obtained by SARSA, ACLA, DoubleQ and QV scheduling policies. The highest amount of GBR moderate rewards is obtained when the QVMAX and QVMAX2 policies are performed revealing the impossibility of these policies to converge to the terminal controller state when $\mathcal{RWI}_t^{FGDP} = 1$. For this reason, QVMAX and QVMAX2 procedures provide the worst performance when the GBR satisfaction levels are measured. Even if the QV policy is not able to assure the satisfactory level of feasible TTIs from the viewpoint of FGDP objectives, the GBR particular objective remains a plausible choice. When the delay reward types are considered (Fig. 7.24), the best percentage of TTIs with the maximum rewards is obtained when SARSA, ACLA, QV2 and Q-L policies are exploited. The percentage of moderate rewards is $\overline{p}_{TTI}^{FGDP,mRW} = 97\%$ for the QV learning which

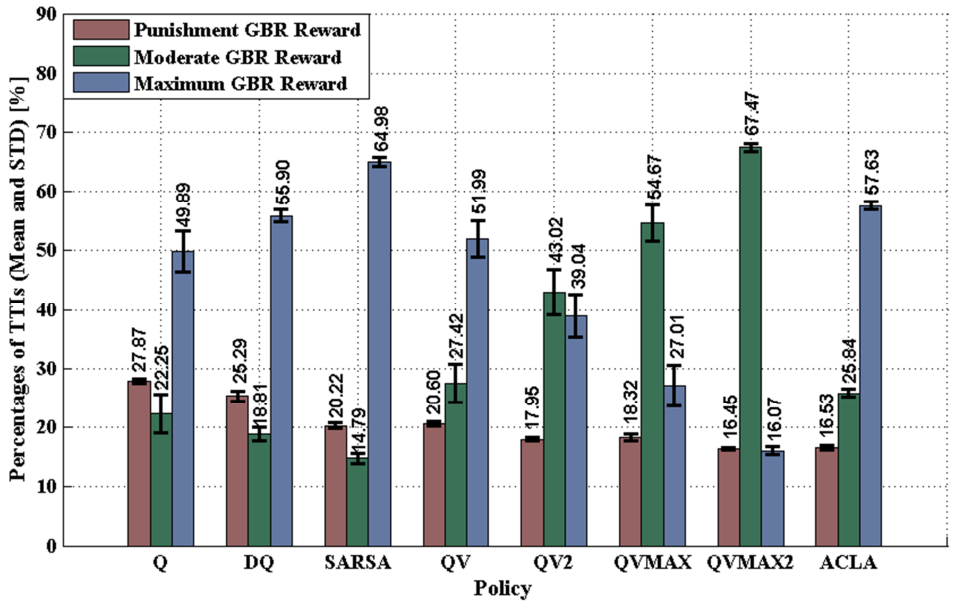


Fig. 7.23 Mean Percentages of TTIs for Punishment, Moderate and Maximum GBR Rewards with the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

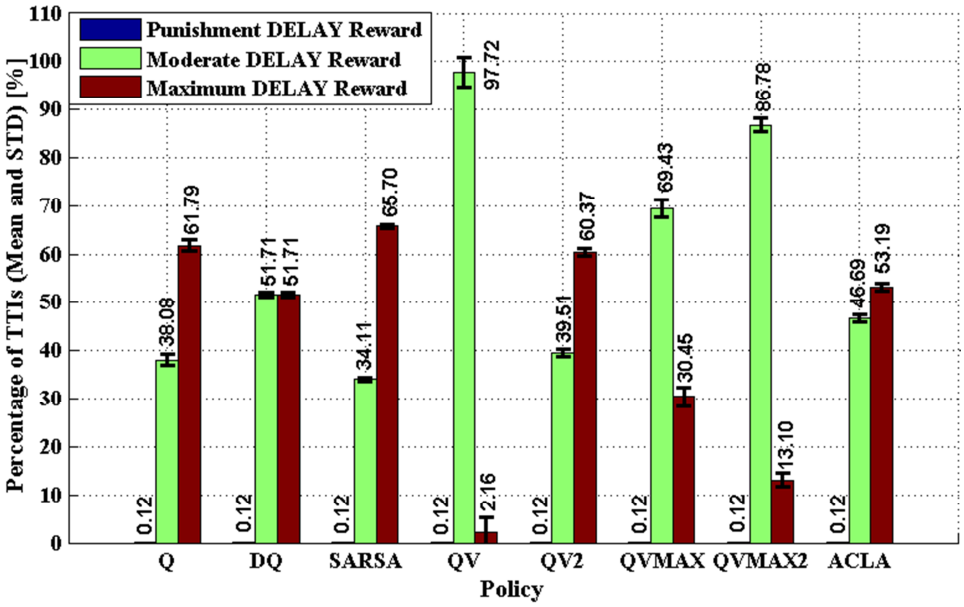


Fig. 7.24 Mean Percentages of TTIs for Punishment, Moderate and Maximum Delay Rewards with the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

implies in fact the degradation of the overall FGDP performance since the HoL delay objective is satisfied only for 2.16% from the entire exploitation time. Figure 7.25 shows the performance of the PDR rewards for the given interval of the windowing factor decided by the CACLA2+ fairness policy. The highest

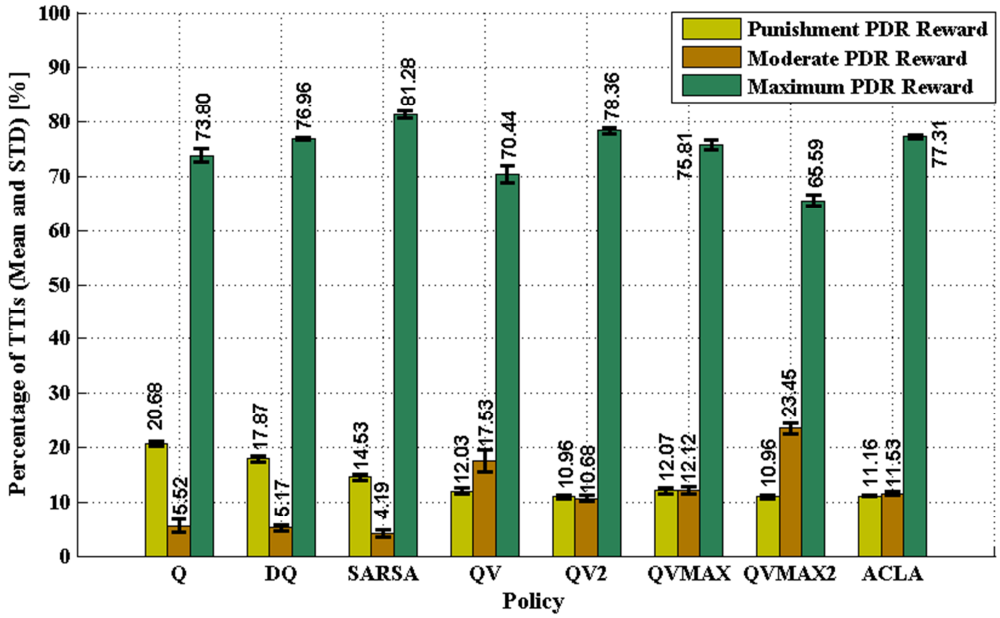


Fig. 7.25 Mean Percentages of TTIs for Punishment, Moderate and Maximum PDR Rewards with the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

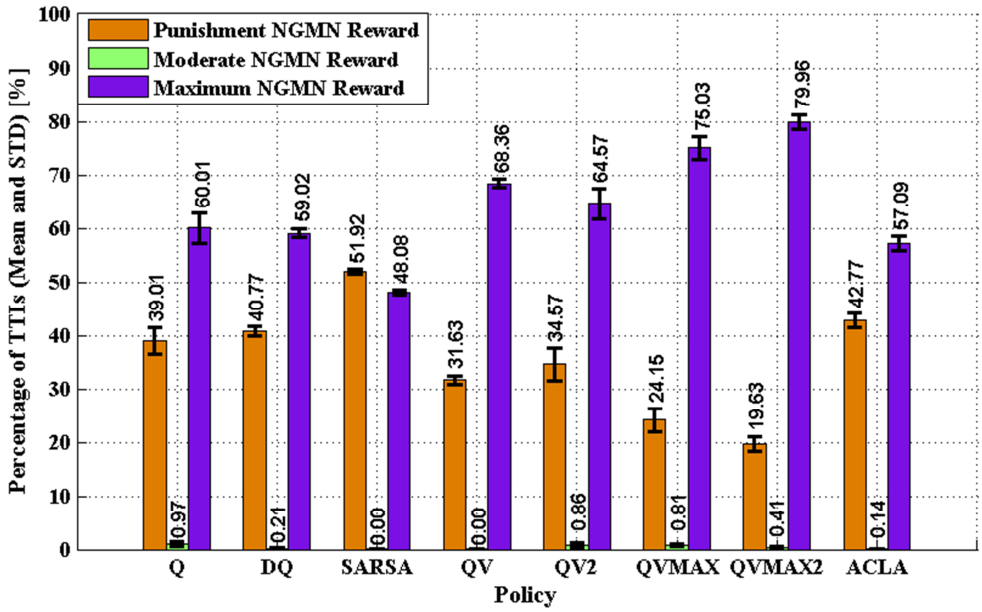


Fig. 7.26 Mean Percentages of TTIs for Punishment, Moderate and Maximum NGMN Fairness Rewards with the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

amount of the PDR maximum rewards in the exploitation stage is obtained through the SARSA, ACLA and QV2 policies. From the NGMN perspective (Fig. 7.26), SARSA achieves the lowest percentage of TTIs when the scheduler is declared feasible from the fairness requirement perspective. Unfortunately, this is

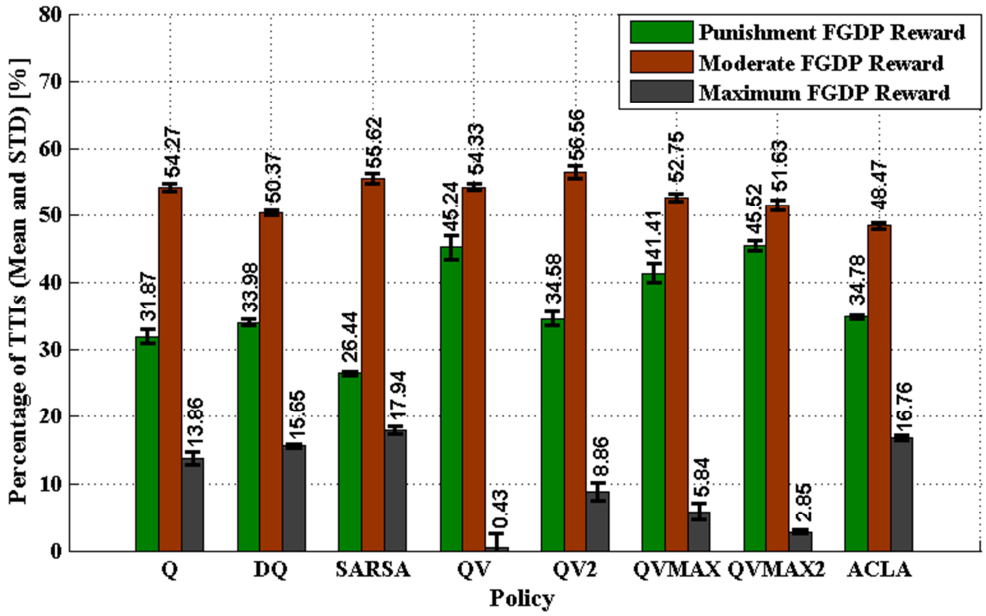


Fig. 7.27 Mean Percentages of TTIs for Punishment, Moderate and Maximum FGDP Rewards with the CBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

the main reason why $\bar{p}_{TTI}^{FGDP,100\%}$ is reduced with more than 40% when compared with $\bar{p}_{TTI}^{GDP,100\%}$ for the exploited SARSA policy. Both QVMAX and QVMAX2 scheduling policies enhance the $\bar{p}_{TTI}^{FGDP,100\%}$ performance when the fairness objective is considered. The QV learning indicates a percentage of the maximum rewards of about $\bar{p}_{TTI}^{F,MRW} = 68\%$, being able to outperform other RL candidates from the viewpoint of fairness and GBR objectives (Fig. 7.19).

By unifying the considered QoS objective reward types from Figs. 7.23, 7.24, 7.25 and 7.26, the multi-objective reward performances are obtained in Fig. 7.27. SARSA, ACLA and DoubleQ learning procedures perform the best in terms of $\bar{p}_{TTI}^{FGDP,MRW}$, whereas the highest amount of punishment rewards is obtained by performing QV, QVMAX and QVMAX2 scheduling policies. QV2 shows an acceptable percentage of TTIs with the maximum rewards when compared with other RL candidates and assures the highest amount of moderate rewards by outperforming at the same time other proposed or existing scheduling methods when the performance domain of $\left[\bar{p}_{TTI}^{FGDP,50\%}, \bar{p}_{TTI}^{FGDP,85\%}\right]$ is considered (Fig. 7.22).

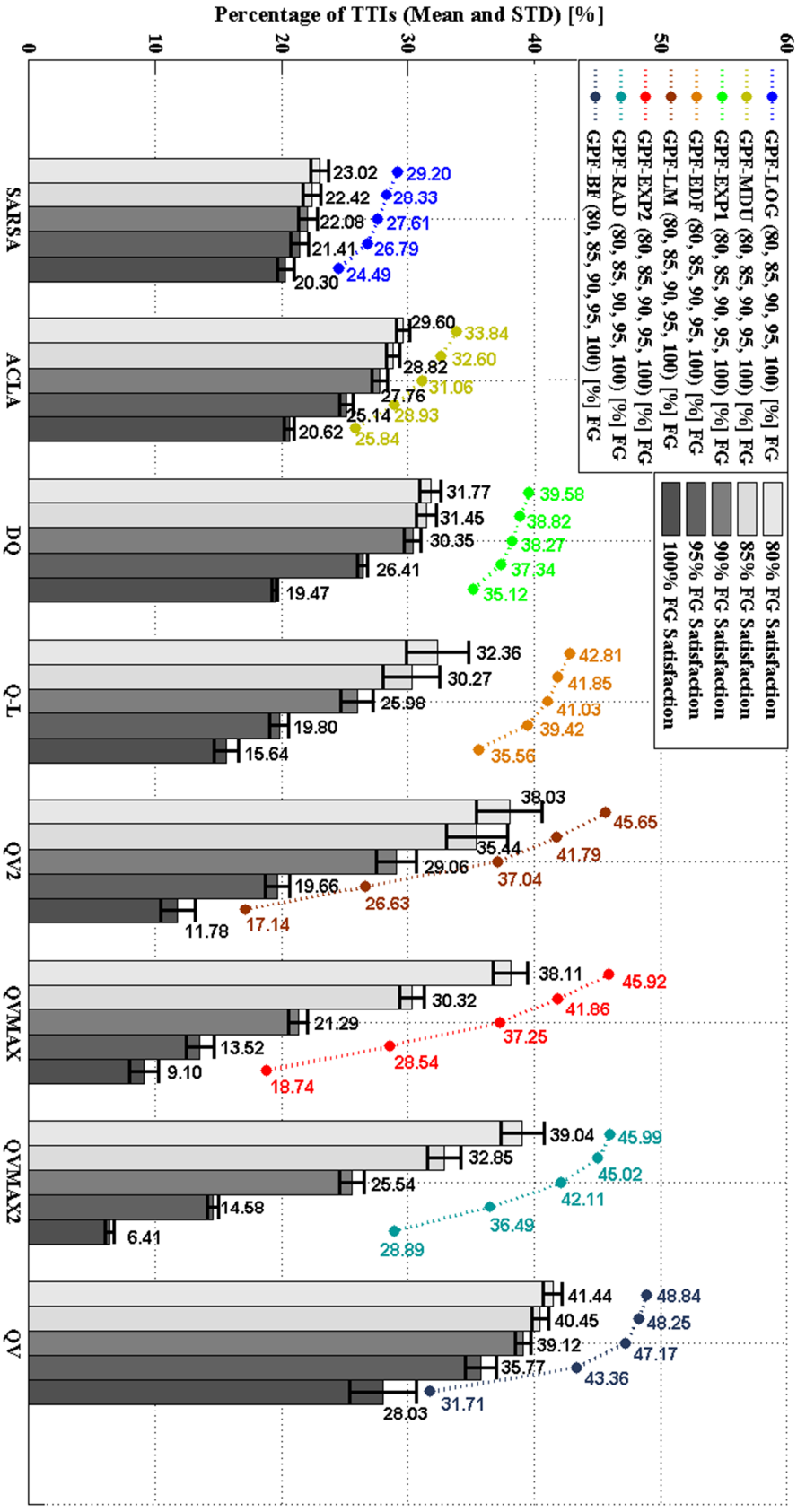


Fig. 7.28 Mean Percentages of TTIs for $\{100\%;95\%;90\%;85\%;80\%\}$ FG Satisfaction with the CBR Traffic Type and a Dynamic Windowing Factor

The mean percentages of TTIs for FG objectives are analyzed in Fig. 7.28 where the considered satisfaction levels are $\left\{\overline{p}_{TTI}^{FG,80\%}, \overline{p}_{TTI}^{FG,85\%}, \overline{p}_{TTI}^{FG,90\%}, \overline{p}_{TTI}^{FG,95\%}, \overline{p}_{TTI}^{FG,100\%}\right\}$. As shown in Fig. 7.28, GPF-MDU and GPF-EXP1 offer the best performance from the percentages of $\overline{p}_{TTI}^{FG,100\%}$ points of view when compared with any other RLs or scheduling rules. When the performance levels of $\left\{\overline{p}_{TTI}^{FG,85\%}, \overline{p}_{TTI}^{FG,90\%}, \overline{p}_{TTI}^{FG,95\%}\right\}$ are analyzed, GPF-BF and GPF-RAD obtain the best results. For the particular case of the percentage of TTIs $\overline{p}_{TTI}^{FG,80\%}$, the GPF-LM and GPF-EXP2 static scheduling rules provide comparable performances when matched against the aforementioned disciplines being oriented on the GBR objective.

When training the QoS MLPNN functions based on the multi-objective reward function $\mathcal{RW}_t^{FGDP}$, the main focus of the learned policy is to increase the number of feasible TTIs from the viewpoint of combined FGDP objectives. This is the explanation of why the learned policies are not able to outperform the existing rules from the perspective of FG multi-objective criterion since the learning target also includes the HoL delay and PDR objectives.

The same satisfaction levels are studied in Fig. 7.29 for the GDP tradeoff performances of different scheduling rules and RL algorithms. The Q-L, DoubleQ, SARSA and ACLA policies outperform other candidates for the entire set of $\left\{\overline{p}_{TTI}^{GDP,80\%}, \overline{p}_{TTI}^{GDP,85\%}, \overline{p}_{TTI}^{GDP,90\%}, \overline{p}_{TTI}^{GDP,95\%}, \overline{p}_{TTI}^{GDP,100\%}\right\}$. As expected, SARSA performs the best from the perspective of the performance set $\left\{\overline{p}_{TTI}^{GDP,90\%}, \overline{p}_{TTI}^{GDP,95\%}, \overline{p}_{TTI}^{GDP,100\%}\right\}$ by indicating a gain of $\{5.06\%, 8.18\%, 16.96\%\}$ when compared with the best scheduling rule GPF-LOG. The DoubleQ policy outperforms SARSA and other scheduling policies from the percentage of $\left\{\overline{p}_{TTI}^{GDP,80\%}, \overline{p}_{TTI}^{GDP,85\%}\right\}$ points of view and indicates a gain set of about $\{5.37\%, 4.66\%\}$ when matched against the GPF-LOG scheduling rule. The QV, QVMAX and QVMAX2 scheduling policies provide an unsatisfactory number of feasible TTIs and high variation factors which require, in fact, more training epochs than other RL approaches in the exploration stage.

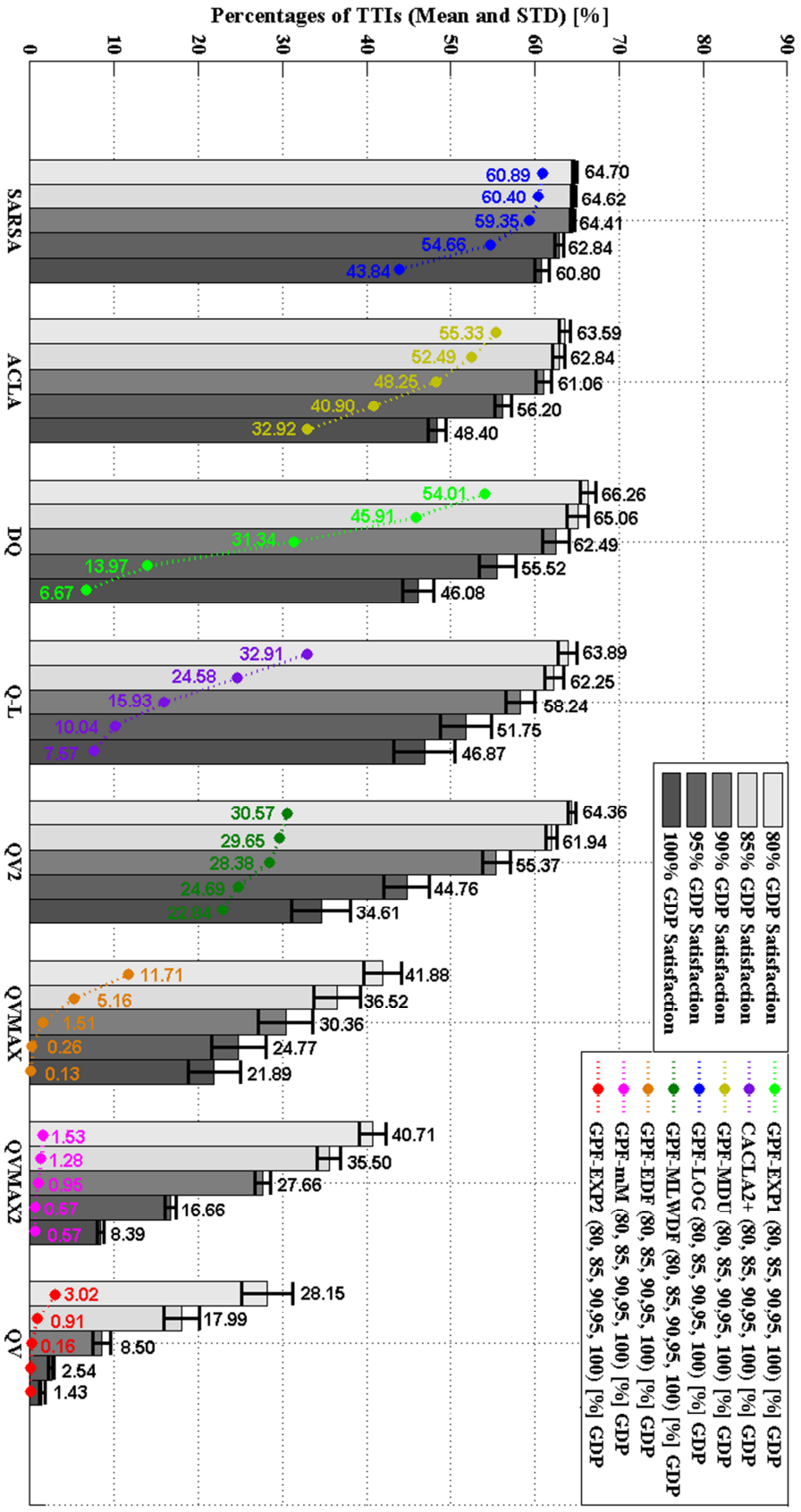


Fig. 7.29 Mean Percentages of TTIs for $\{100\%;95\%;90\%;85\%;80\%\}$ GDP Satisfaction with the CBR Traffic Type and a Dynamic Windowing Factor

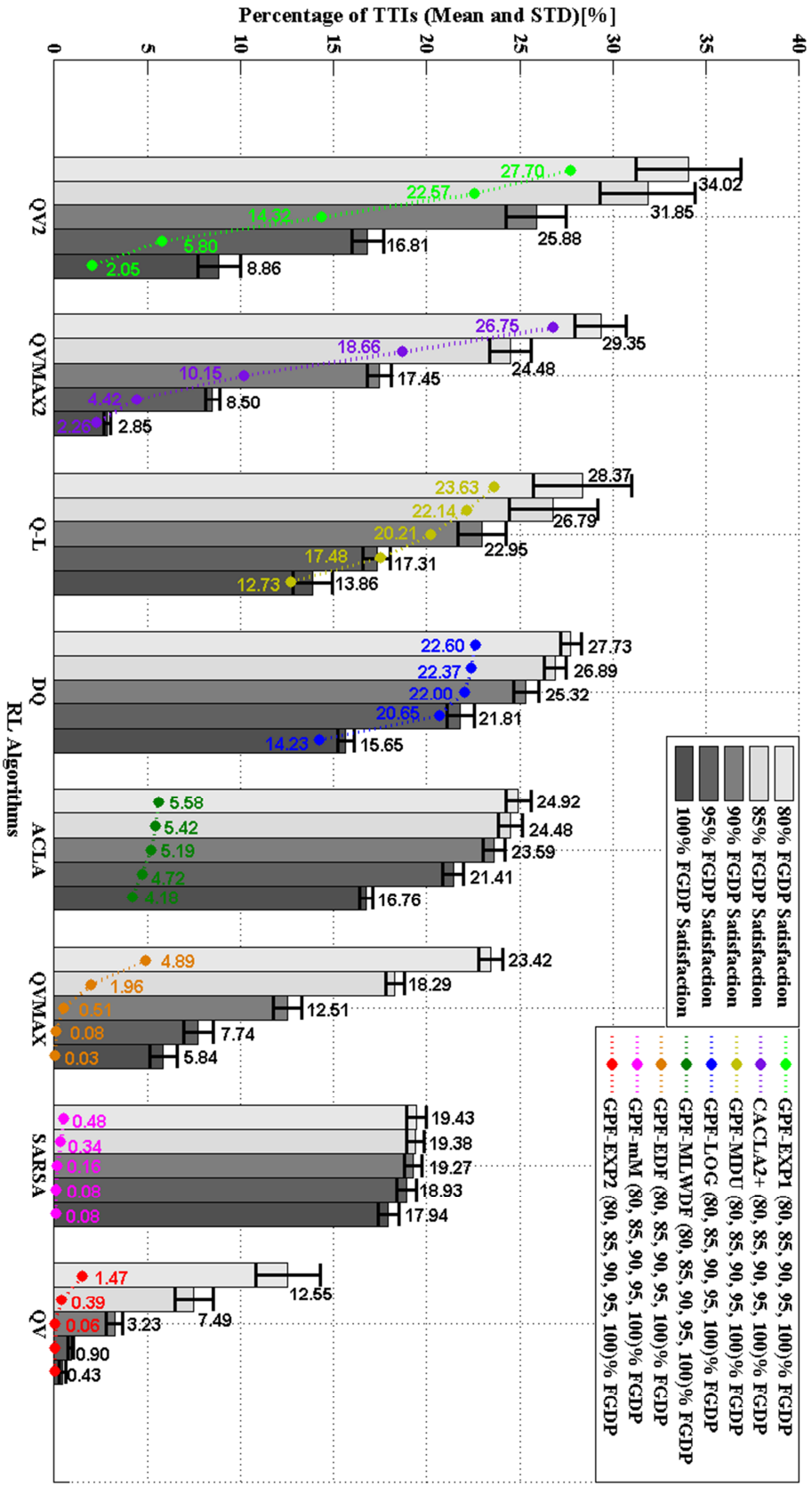


Fig. 7.30 Mean Percentages of TTIs for $\{100\%; 95\%; 90\%; 85\%; 80\%\}$ FGDP Satisfaction with the CBR Traffic Type and a Dynamic Windowing Factor

When the NGMN objective is considered together with other analyzed objectives (Fig. 7.30), the SARSA scheduling policy is the best choice only from the $\bar{p}_{TTI}^{FGDP,100\%}$ perspective by showing a percentage gain of 3.71% when compared with the GPF-LOG rule. When the percentages $\left\{\bar{p}_{TTI}^{FGDP,90\%}, \bar{p}_{TTI}^{FGDP,95\%}\right\}$ are measured, the DoubleQ procedure outperforms GPF-LOG by about $\{3.32,1.1\}\%$. For lower FGDP satisfaction levels such as $\left\{\bar{p}_{TTI}^{FGDP,80\%}, \bar{p}_{TTI}^{FGDP,85\%}\right\}$, the QV2 policy obtains the percentage gain set of $\{6.32,9.28\}\%$ when matched against the GPF-EXP1 scheduling rule. By introducing the NGMN objective, the overall FGDP satisfaction level is decreased. The most affected is the SARSA policy which loses more than 44% for the considered percentage domain due to the difficulty in stabilizing the scheduling policy in the NGMN feasible region (Fig. 7.26). Despite of these aspects, the SARSA scheduling policy for the CBR traffic type is sustainable due to the fact that mean percentage of FGDP feasible TTIs is maximized when matched against other RL and static scheduling rule approaches, the STD values for the entire performance domain are minimized and the punishment rewards register the lowest amount when the exploited SARSA policy is compared against other RL scheduling policies.

7.3.5.3 DSR-CMOO Focusing on HoL Delay, PDR, GBR and NGMN Fairness Objectives with the VBR Traffic Type

In the VBR traffic type case, the GBR oriented scheduling rules assure the highest percentage of TTIs for the satisfied bearers when the single GBR objective is considered (Fig. 7.31) in contrast with the CBR traffic type case where the GPF-LOG and GPF-MDU outperform GPF-BF and GPF-RAD scheduling disciplines. The QV scheduling policy indicates the best performance from the viewpoint of GBR satisfaction levels. Other scheduling policies degrade the GBR performance starting from the percentage of $\bar{p}_{TTI}^{G,80\%}$. When the tradeoff between the NGMN fairness and GBR objectives is considered (Fig. 7.32), the GPF-LM, GPF-BF and GPF-RAD scheduling rules are able to outperform other

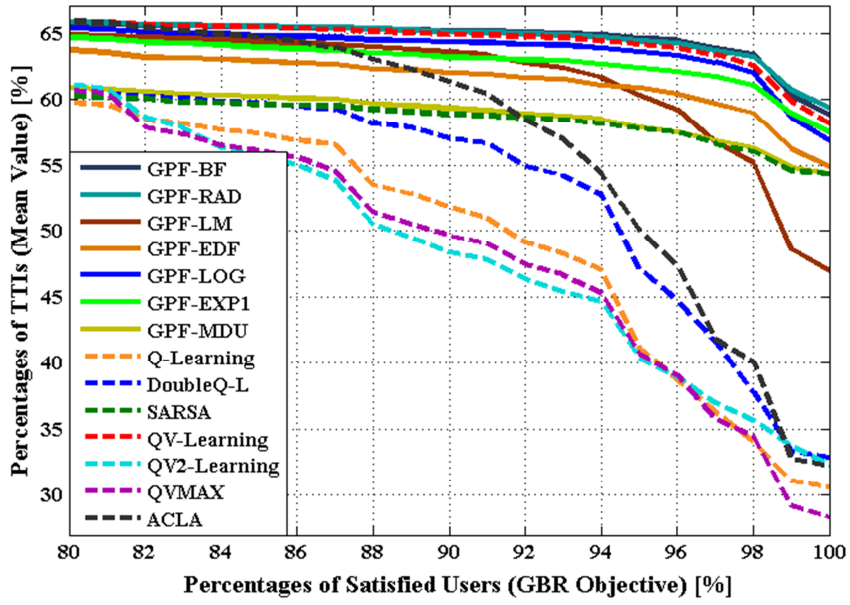


Fig. 7.31 Mean Percentages of TTIs vs. Percentages of GBR Satisfied Bearers for the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

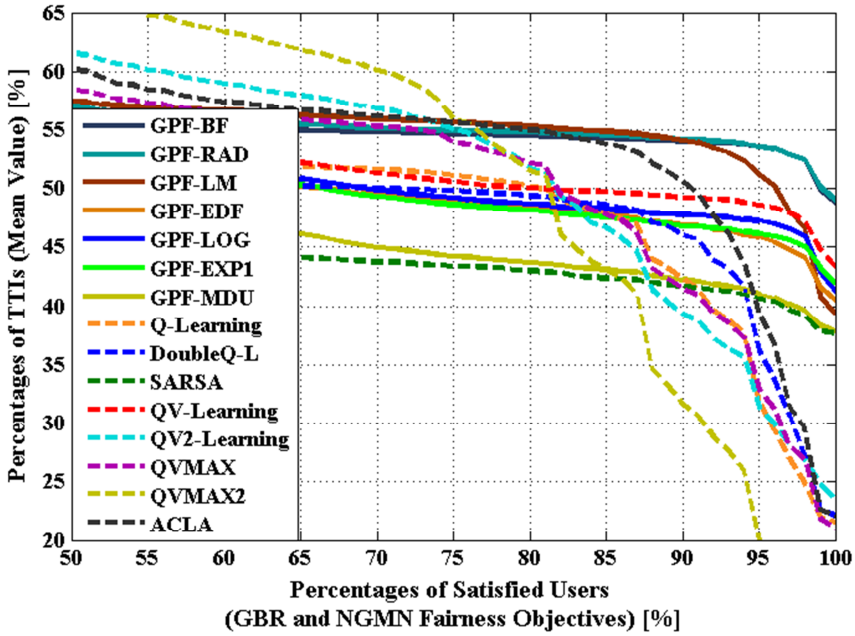


Fig. 7.32 Mean Percentages of TTIs vs. Percentages of FG Satisfied Bearers for the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

candidates for the performance domain of $\left[\overline{p}_{TTI}^{FG,75\%}, \overline{p}_{TTI}^{FG,100\%}\right]$. For the performance domain lower than $\overline{p}_{TTI}^{G,80\%}$, ACLA offers comparable results when compared with GBR oriented disciplines. For the entire satisfaction domain of FG objectives, the

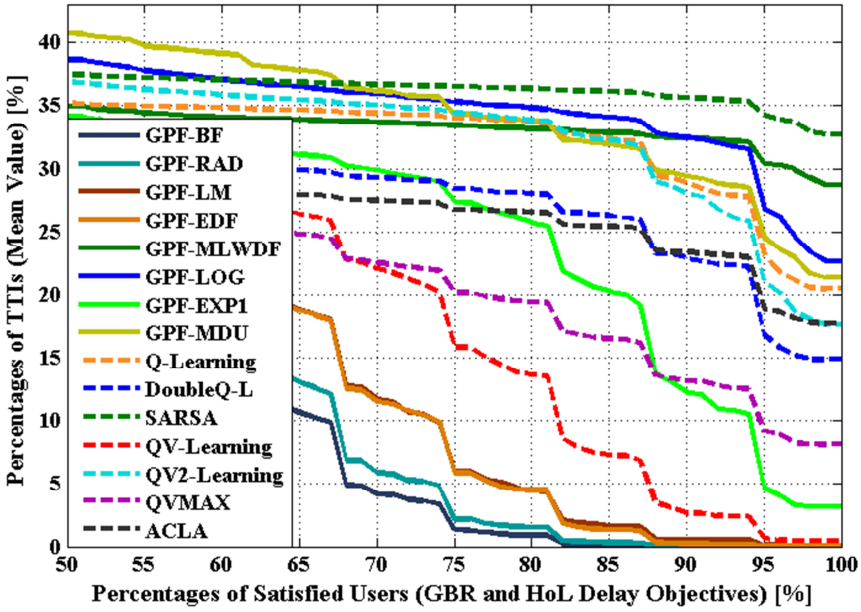


Fig. 7.33 Mean Percentages of TTIs vs. Percentages of GD Satisfied Bearers for the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

QV learning shows the best tradeoff between the scheduling rules oriented on the GBR and HoL delay objectives. Other policies decrease in performance when exceeding the performance percentage threshold of FG feasible TTIs $\bar{p}_{TTI}^{FG,75\%}$.

When the GD multi-objective criterion is analyzed, the GBR scheduling rules show the worst performance as indicated in Fig. 7.33 while GPF-MLWDF and GPF-LOG perform better than other scheduling rules for the interval of $\left[\bar{p}_{TTI}^{GD,75\%}, \bar{p}_{TTI}^{GD,100\%}\right]$. The SARSA policy shows the best performance when the domain of $\left[\bar{p}_{TTI}^{GD,75\%}, \bar{p}_{TTI}^{GD,100\%}\right]$ is analyzed. By considering the HoL delay objective, the QV policy shows the worst performance when compared with other RL approaches since it is focused more on the GBR and NGMN fairness objectives.

The SARSA, QV2 and DoubleQ scheduling policies outperform the GPF-LOG and GPF-MDU rules from the viewpoint of the tradeoff among the DP objectives (Fig. 7.34). In fact, the SARSA policy gains more than 10% of DP feasible TTIs when compared against other scheduling rules. Each scheduling rule works better under a given number of bearers and under different profiles of QoS requirements. By switching from one QoS requirement profile to another at each

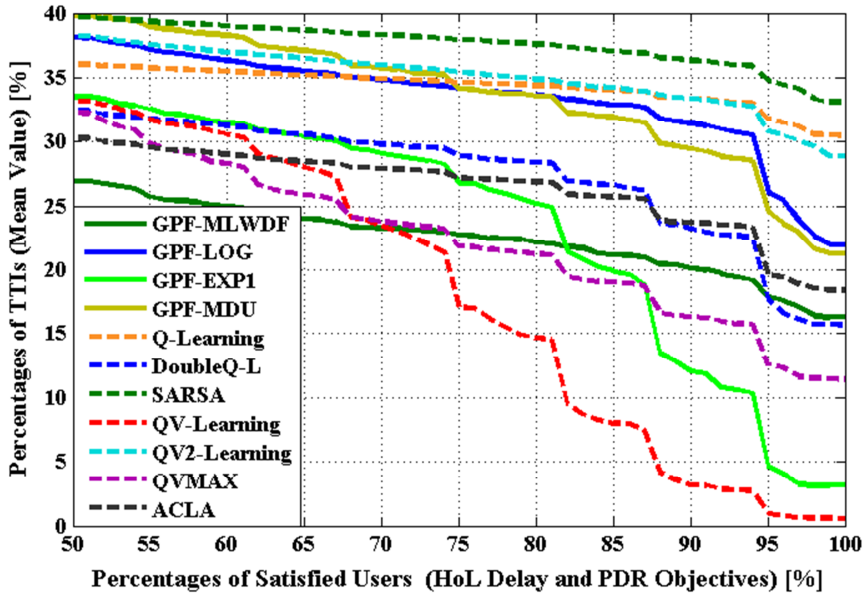


Fig. 7.34 Mean Percentages of TTIs vs. Percentages of DP Satisfied Bearers for the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

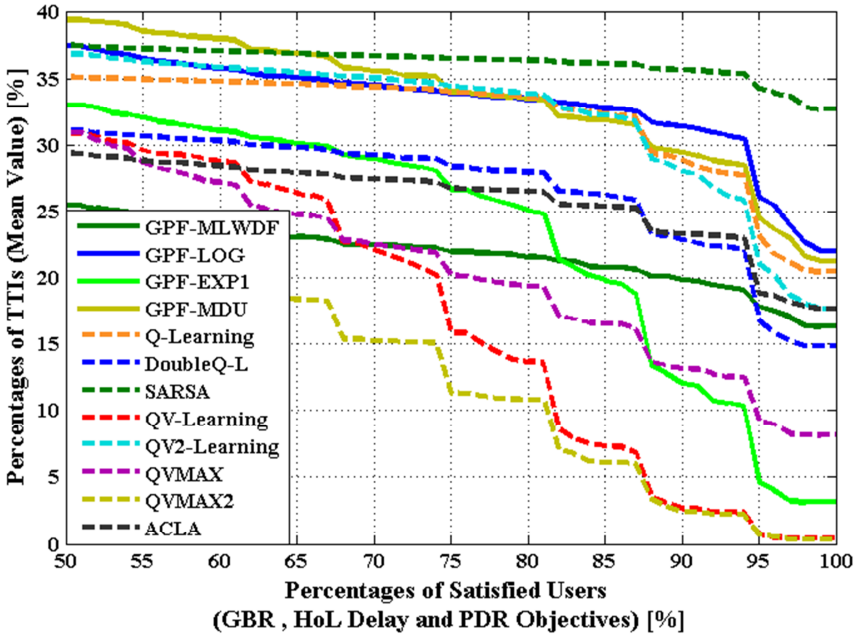


Fig. 7.35 Mean Percentages of TTIs vs. Percentages of GDP Satisfied Bearers for the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

1000 TTIs, SARSA is able to find the optimal set of rules which can enhance the convergence in the feasible state. In particular, when GPF-LOG, GPF-MDU, GPF-MLWDF and GPF-EXP1 are combined in a very intelligent manner based on the given traffic and QoS conditions, the SARSA policy is able to increase the

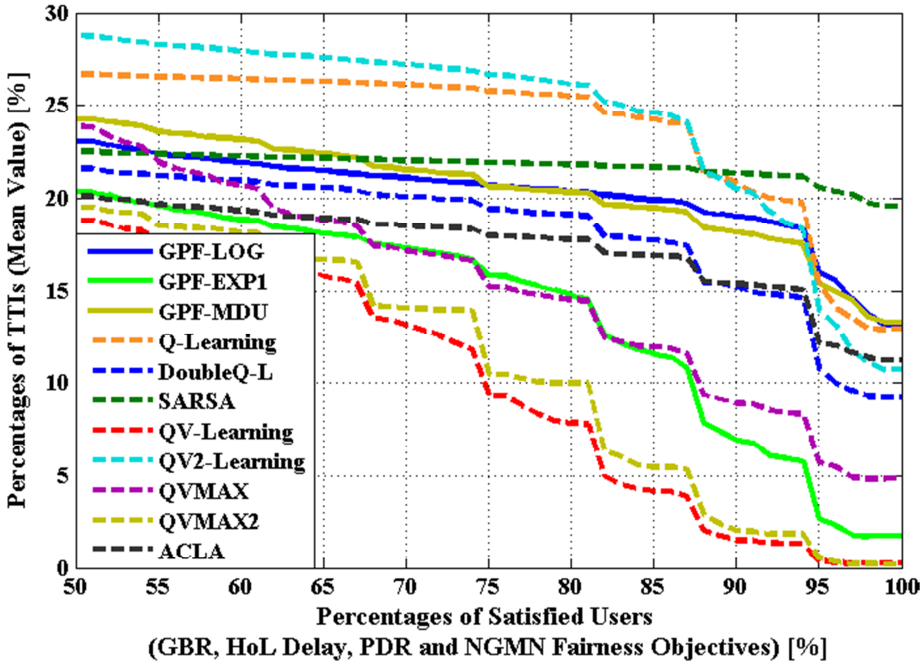


Fig. 7.36 Mean Percentages of TTIs vs. Percentages of FGDP Satisfied Bearers for the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

mean percentage of TTIs for the entire interest domain of DP objectives. By inserting the GBR performance in the DP multi-objective evaluation (Fig. 7.35), SARSA maintains the same gain of GDP feasible TTIs when matched against GPF-LOG and GPF-MDU. The performances of $\overline{p}_{TTI}^{DP,100\%}$ and $\overline{p}_{TTI}^{GDP,100\%}$ remain similar when the SARSA scheduling policy is performed. Only QVMAX2 policy degrades its performance when the GBR objective is inserted in the DP multi-objective evaluation. The NGMN fairness requirement decreases the number of feasible TTIs for SARSA, when compared with the same algorithms, by about 5% (Fig. 7.36). In fact, SARSA performs the best for a more restrictive domain such as $\left[\overline{p}_{TTI}^{FGDP,87\%}, \overline{p}_{TTI}^{FGDP,100\%}\right]$. Beyond of this interval, the QV2 and Q-L policies are the best choices from the FGDP multi-objective tradeoff point of view.

From the GBR objective perspective, SARSA and QV receive the highest amount of maximum rewards in the exploitation period as shown in Fig. 7.37. On the other hand, ACLA, QV2, QVMAX and QVMAX2 indicate the largest amount of moderate rewards which hampers the convergence of the learned policy to the GBR feasible state. When the Q-L policy is performed, the scheduler punishes the

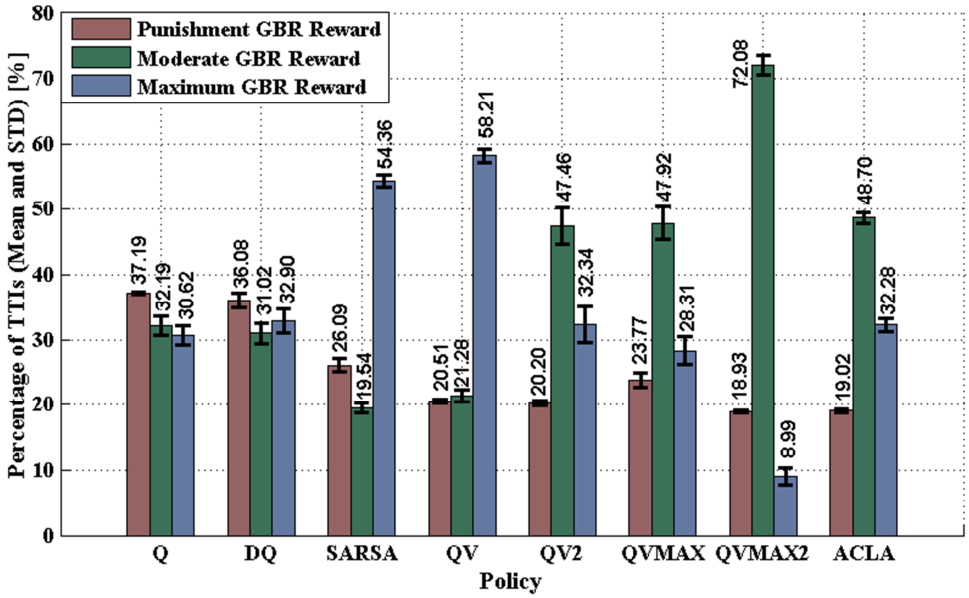


Fig. 7.37 Mean Percentages of TTIs for Punishment, Moderate and Maximum GBR Rewards with the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

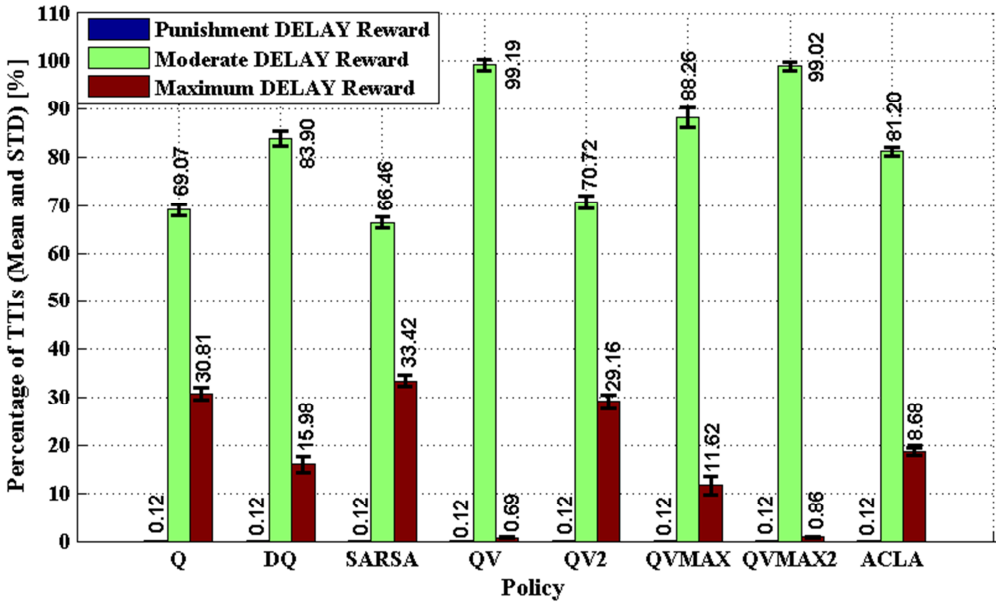


Fig. 7.38 Mean Percentages of TTIs for Punishment, Moderate and Maximum Delay Rewards with the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

applied rules more often in this policy than in any other cases by achieving a level of $\bar{p}_{TTI}^{G,PSH} = 37.19\%$ at the end of the exploitation stage. The QV learning achieves the lowest percentage of punishments when compared with SARSA. For this reason, the QV policy is able to achieve a gain of about 7% from the $\bar{p}_{TTI}^{G,100\%}$ point

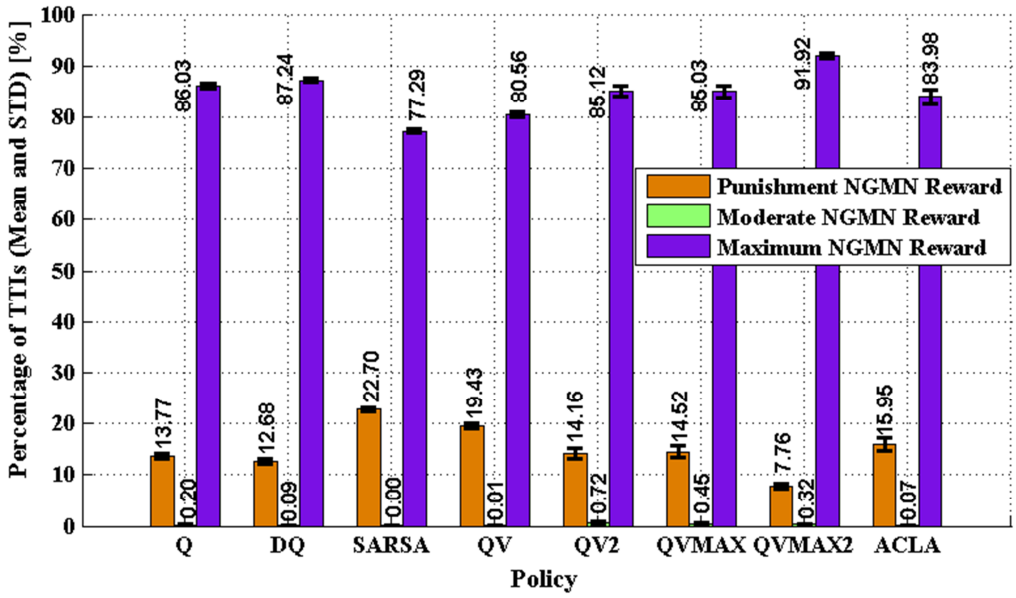


Fig. 7.39 Mean Percentages of TTIs for Punishment, Moderate and Maximum NGMN Fairness Rewards with VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

of view. When the delay reward performance is considered (Fig. 7.38), QV, QVMAX and QVMAX2 denote the highest percentage of TTIs with moderate rewards revealing the degraded performance of these policies when the particular HoL delay reward is taken into account. SARSA and Q-L perform the best while obtaining the levels of $\overline{p}_{TTI}^{D,MRW} = 33.42\%$ and $\overline{p}_{TTI}^{D,MRW} = 30.81\%$, respectively. Figure 7.39 shows the performance of each scheduling policy in terms of the NGMN fairness requirement. Similar to the CBR case, the SARSA policy provides the worst performance by indicating a percentage of TTIs with the maximum rewards of about 77.29% by maximizing at the same time the number of punishments received in the exploitation period. Based on these reasons, the overall multi-objective performance is degraded when the fairness objective is considered in the multi-objective optimization. The QVMAX2 and DoubleQ learning procedures are considered to be the best choices from the NGMN fairness requirement point of view. SARSA aims to minimize the windowing factor by harming the fairness performance on one hand and aims to increase the percentage of the maximum rewards for the PDR objective on the other hand (Fig. 7.40). The Q-L policy assures the highest percentage of punishments for the PDR objective due to the fact that it uses a lower windowing factor, when compared

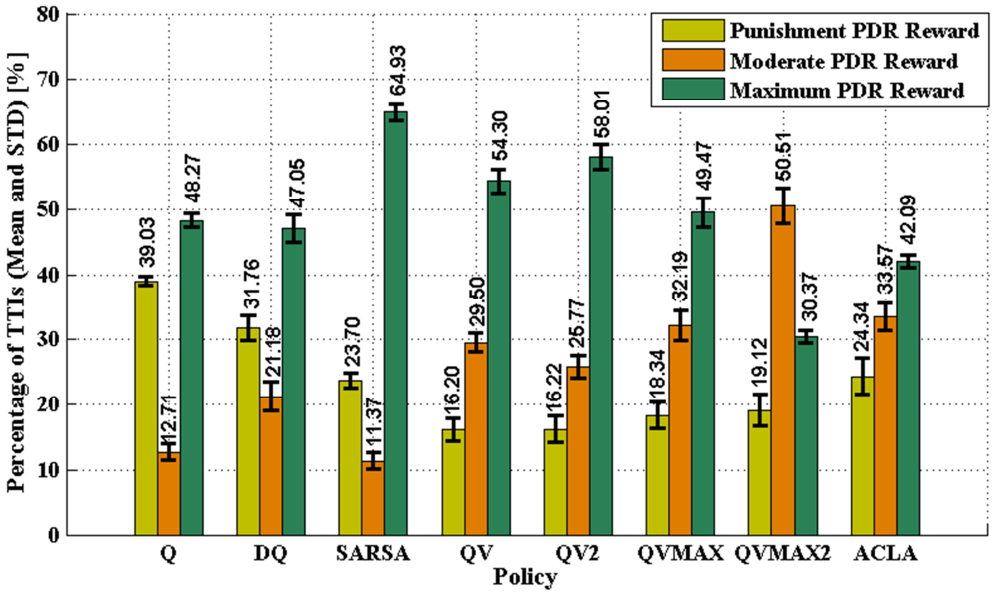


Fig. 7.40 Mean Percentages of TTIs for Punishment, Moderate and Maximum PDR Rewards with the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

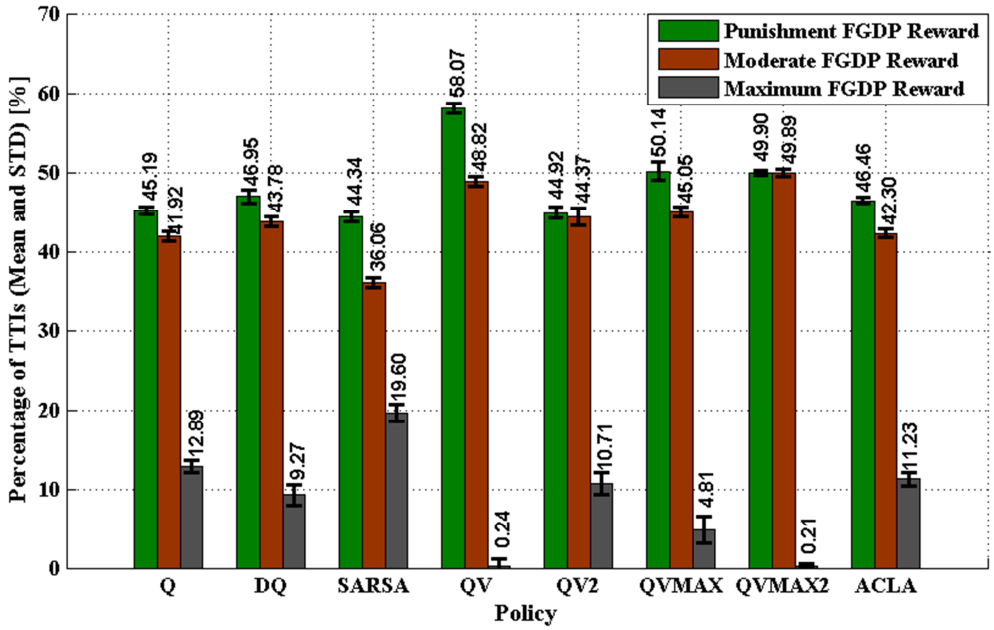


Fig. 7.41 Mean Percentages of TTIs for Punishment, Moderate and Maximum FGDP Rewards with the VBR Traffic Type and a Dynamic Windowing Factor of $\rho \in [2.5;50]$

with other candidates, which in fact implies unstable rewards. The reason is also viable for the GBR objective, where Q-L policy receives the largest amount of punishments (Fig. 7.37). For the PDR objective, QVMAX2 shows an increased level of moderate rewards since it uses larger windowing factors and the impact of

the immediate effect of the applied scheduling rule is not sensed in the instantaneous reward value. The performance of the multi-objective reward which considers the difference between the consecutive intrinsic rewards is analyzed in Fig. 7.41. SARSA obtains the highest amount of the maximum rewards of about $\bar{p}_{TTI}^{FGDP,MRW} = 19.6\%$. Due to the higher maximum rewards for the HoL and NGMN objectives, the Q-L scheduling policy is able to achieve a mean percentage level of $\bar{p}_{TTI}^{FGDP,MRW} = 12.89\%$, being considered the second best choice from the FGDP tradeoff point of view. Similar to the CBR traffic type, the Pareto arrival bit rate does not bring any improvement in the QV policy which obtains the worst performance when the considered DSR-CMOO problem is performed.

Figure 7.42 presents different performance levels in terms of $\left\{\bar{p}_{TTI}^{FG,80\%}, \bar{p}_{TTI}^{FG,85\%}, \bar{p}_{TTI}^{FG,90\%}, \bar{p}_{TTI}^{FG,95\%}, \bar{p}_{TTI}^{FG,100\%}\right\}$ for the tradeoff between the GBR and NGMN fairness objectives in the cases of the best ranked scheduling rules and RL scheduling policies. As mentioned, GPF-BF and GPF-RAD perform the best from the perspective of $\left\{\bar{p}_{TTI}^{FG,95\%}, \bar{p}_{TTI}^{FG,100\%}\right\}$ by indicating a gain of $\{5.7, 4.94\}\%$ when compared against the best scheduling policy being obtained when exploring with the QV algorithm. For the discrete performance levels of $\left\{\bar{p}_{TTI}^{FG,85\%}, \bar{p}_{TTI}^{FG,90\%}\right\}$, the GPF-EXP2 scheduling rule outperforms any other rule or RL policies. GPF-LM and GPF-EXP2 provide a similar performance in terms of the 80% FG satisfied bearers by indicating a performance level of $\bar{p}_{TTI}^{FG,80\%} = 55.3\%$. Other scheduling rules such as GPF-OPLF, GPF-PLF, GPF-EXP1 and GPF-LOG are not able to outperform the QV policy for the considered FG tradeoff levels.

When the GDP multi-objective performance is studied (Fig. 7.43), SARSA performs the best for the first three levels of the GDP performance such as $\left\{\bar{p}_{TTI}^{GDP,90\%}, \bar{p}_{TTI}^{GDP,95\%}, \bar{p}_{TTI}^{GDP,100\%}\right\}$ by achieving a gain set of $\{5.06, 8.18, 16.96\}\%$ when compared with the GPF-LOG scheduling rule. From the viewpoint of the mean percentage of TTIs when the following performance levels are considered

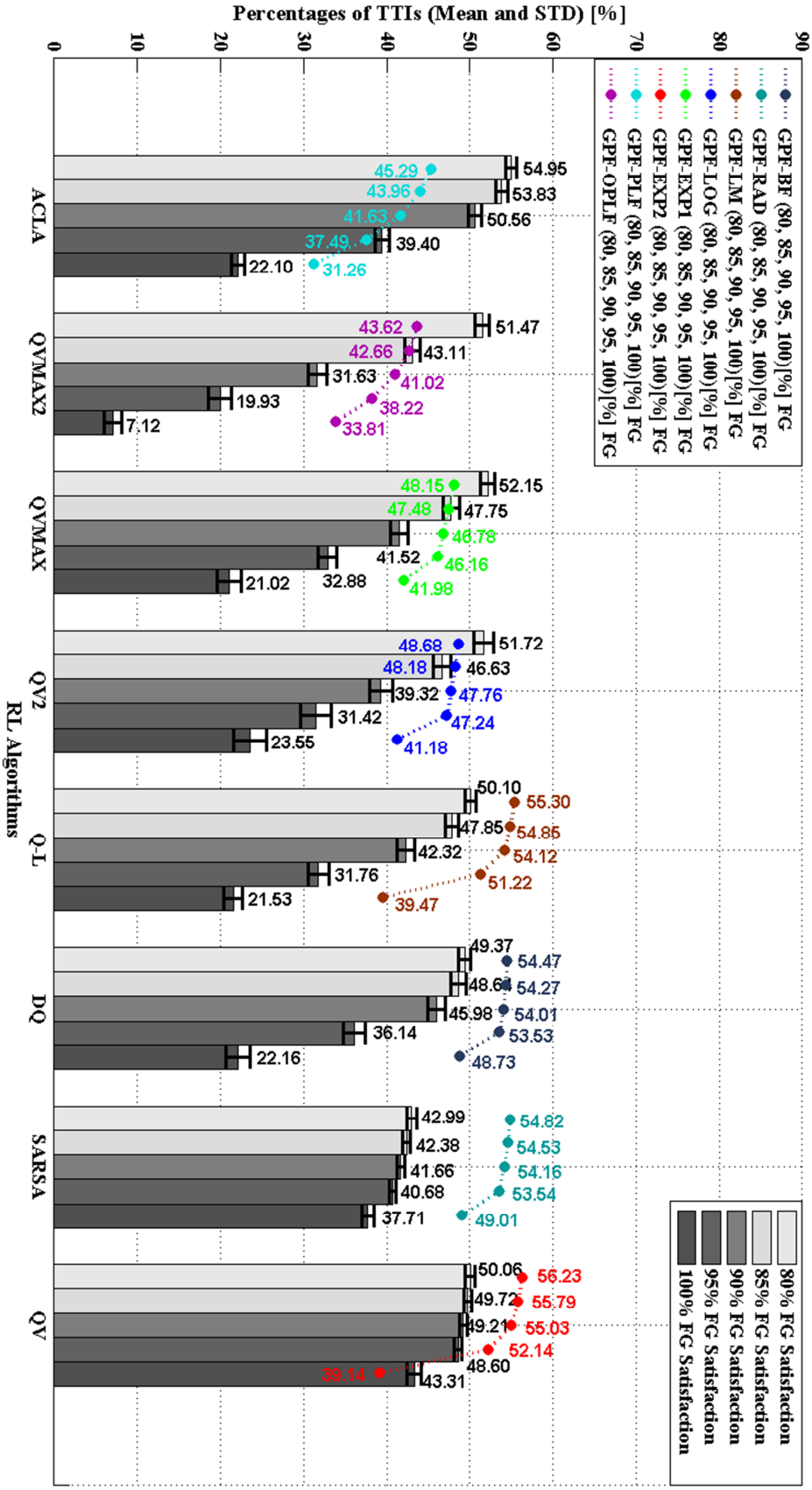


Fig. 7.42 Mean Percentages of TTIs for $\{100\%;95\%;90\%;85\%;80\%\}$ FG Satisfaction with the VBR Traffic Type and a Dynamic Windowing Factor

$\left\{\overline{p}_{TTI}^{GDP,80\%}, \overline{p}_{TTI}^{GDP,85\%}\right\}$, DoubleQ outperforms the main candidate GPF-LOG rule by about $\{4.36, 4.66\}\%$. The highest variations of the plotted results are obtained when performing QV and QVMAX scheduling policies due to the much larger windowing factor involved in the DSR-CMOO optimization problem.

By inserting the NGMN fairness requirement in the multi-objective evaluation levels (Fig. 7.44), the variations of the obtained percentage of TTIs become higher but in acceptable limits without affecting the comparison conclusions for the obtained scheduling policies. As expected, the highest variations are achieved by applying the QVMAX scheduling policy. These variations are caused mainly by the dynamic windowing factor which is changed periodically by the CACLA2+ policy based on the controller state information. When the windowing factor is large, the percentage of moderate rewards for the GBR, NGMN fairness and PDR objectives becomes higher, and implicitly, the multi-objective tradeoff satisfaction levels present higher STD factors. In the QV policy case (Fig. 7.44), the higher variation factors for each percentage level are caused by the fact that the windowing factor is reduced, and implicitly, the immediate reward values start to fluctuate when selecting scheduling rules TTI-by-TTI. Alongside these aspects, SARSA assures lower variation and the best results from the perspective of $\left\{\overline{p}_{TTI}^{FGDP,90\%}, \overline{p}_{TTI}^{FGDP,95\%}, \overline{p}_{TTI}^{FGDP,100\%}\right\}$ performance levels by achieving a gain set of $\{2.36, 4.58, 6.6\}\%$ when compared with GPF-LOG or GPF-MDU scheduling disciplines. With lower STD factors when compared with the Q-L learning algorithm, the QV2 scheduling policy achieves the highest amount of TTIs for the performance levels of $\left\{\overline{p}_{TTI}^{FGDP,80\%}, \overline{p}_{TTI}^{FGDP,85\%}\right\}$ when the FGDP multi-objective tradeoff evaluation is considered.

Being able to maximize the mean percentage of FGDP feasible TTIs $\overline{p}_{TTI}^{FGDP,100\%}$ with the lowest variation of the results when the SARSA QoS agent and the CACLA2+ fairness agent are performed, the obtained scheduling policy is sustainable for the VBR traffic type. At the same time, the amount of punishments is minimized when the FGDP multi-objective criterion is considered (Fig. 7.41).

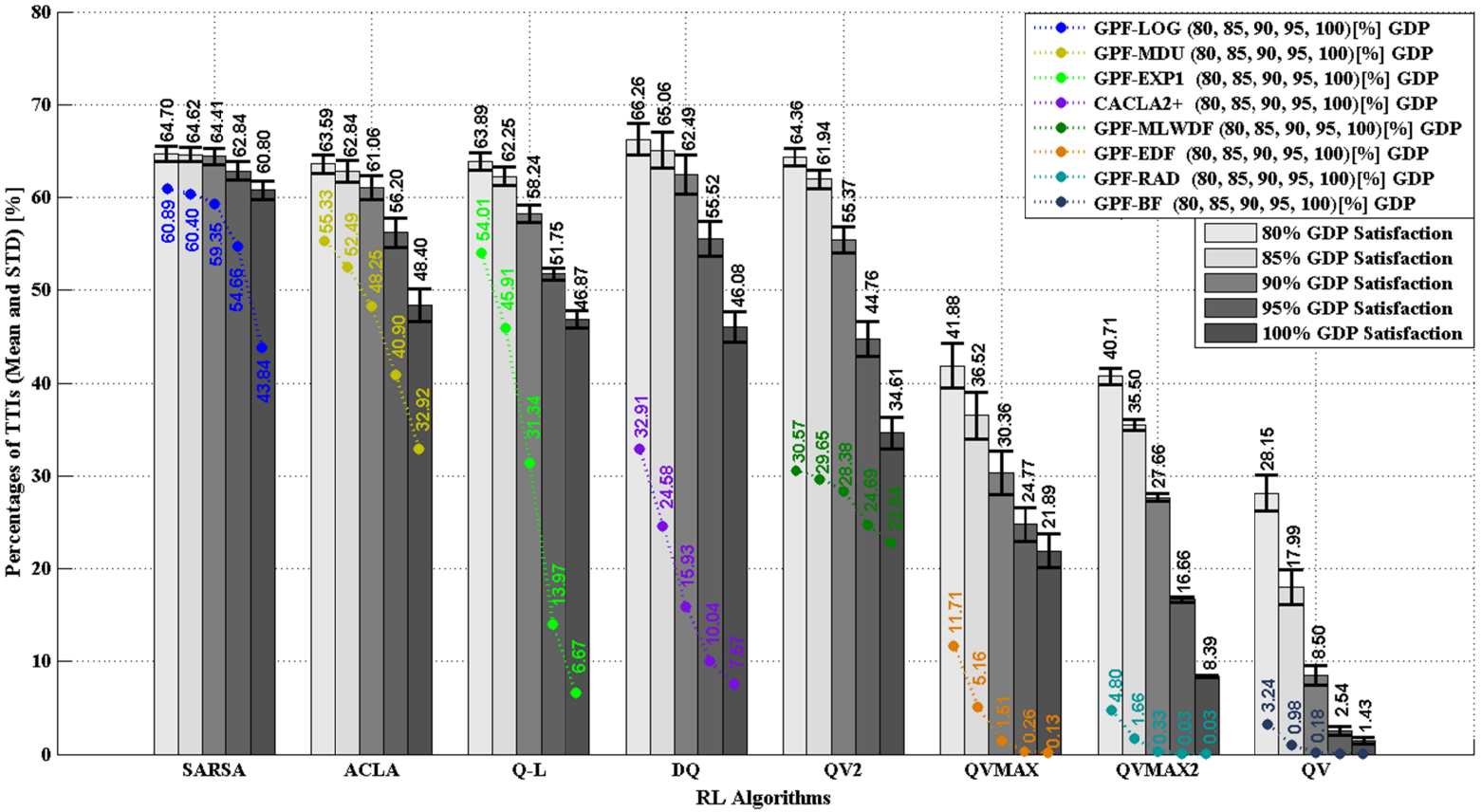


Fig. 7.43 Mean Percentages of TTIs for $\{100\%;95\%;90\%;85\%;80\%\}$ GDP Satisfaction with the VBR Traffic Type and a Dynamic Windowing Factor

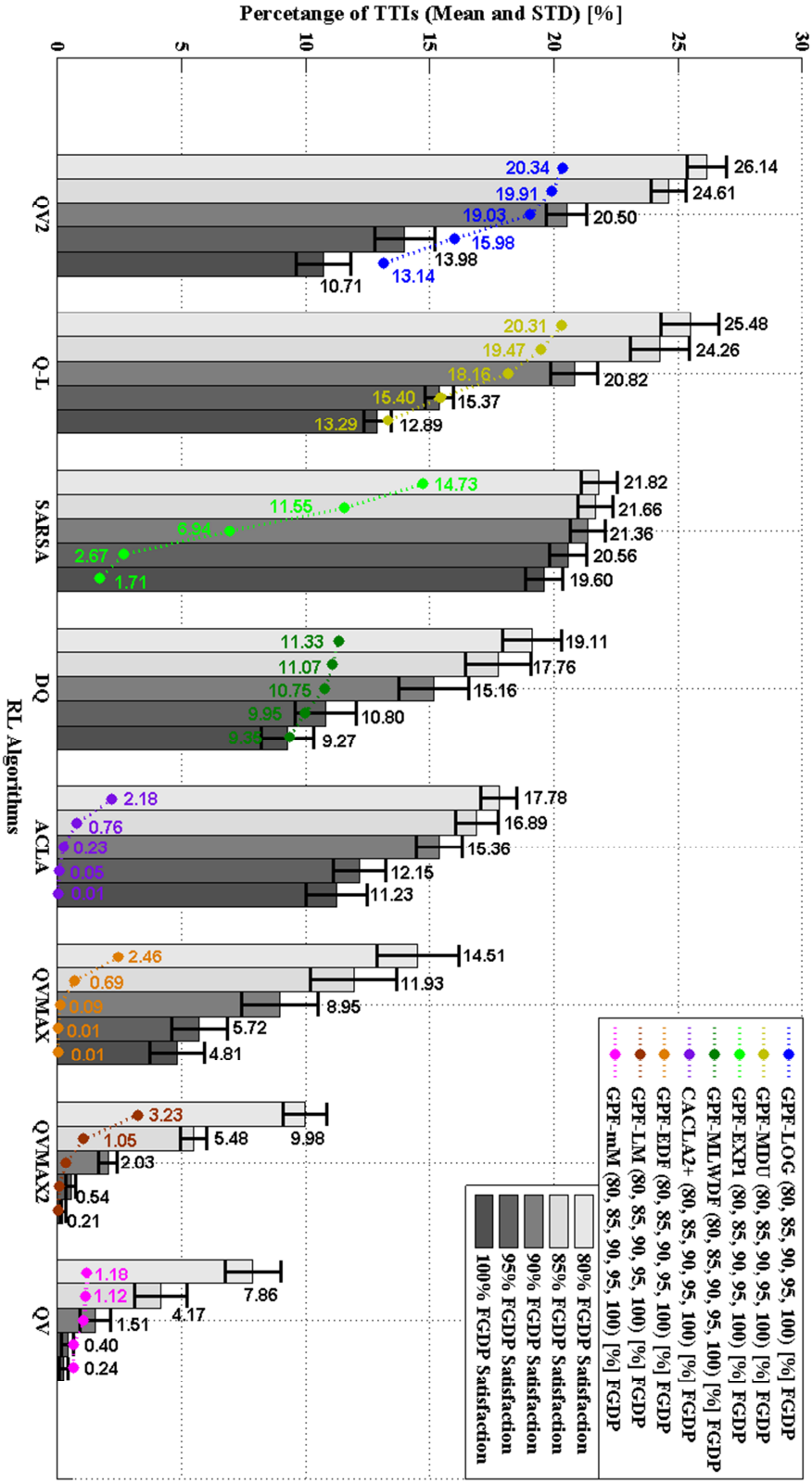


Fig. 7.44 Mean Percentages of TTIs for {100%;95%;90%;85%;80%} FGDP Satisfaction with the VBR Traffic Type and a Dynamic Windowing Factor

7.4 Summary

Two types of concurrent optimization problems have been analyzed in this chapter in terms of the DP and FGDP multi-objective tradeoff evaluation for both CBR and VBR traffic types. When the DP DSR-CMOO is performed, there are two aspects with major impacts on the performance of scheduling policies: the windowing factor which is involved in the PDR objective evaluation and the particularities of delay based scheduling rules in terms of the standard deviation of HoL delays. It is proven with eligible results that the GPF-LOG rule performs the best when compared with other existing schemes by minimizing, at the same time, the mean and STD parameters for the HoL packet delay observations. In this sense, the percentage of TTIs for 100% DP satisfied bearers increases when compared with other scheduling rules. By combining all types of scheduling rules focusing on delay and PDR objectives, the proposed scheduling policies are able to outperform any other singular scheduling rule from the viewpoint of the mean percentage of DP feasible TTIs $\left(\overline{p}_{TTI}^{DP,100\%}\right)$. When the windowing factor increases, the proposed policies are not able to apply the optimal actions since the reward function is not able to sense the immediate impact of each applied scheduling rule. In this sense, for lower satisfaction domains, the GPF-LOG rule indicates better performances. However, the learned policies perform much better than the standard scheduling rules from the viewpoint of the mean percentage of feasible TTIs when the 100% DP satisfaction level is considered for both types of traffic.

In order to find the optimum windowing factor for the evaluation of the NGMN, GBR and PDR objectives, the CACLA2+ RL approach is proposed to adapt a third action in terms of the windowing factor step. When the FGDP DSR-CMOO MDP combinatorial problems are considered, CACLA2+ adapts the fairness parameters and the windowing factor at each time instant when the GPF-DP scheduling rule is selected by the QoS controller. In this sense, the proposed SARSA scheduling policy achieves a gain greater than 7% when compared with the best ranked scheduling policies from the FGDP multi-objective tradeoff point of view. For lower FGDP satisfaction levels, a better performance can be obtained by exploiting the QV2 scheduling policies for both traffic types.

Chapter 8

Conclusions

8.1 Chapter Outline

The motivation of using the DSR-SMOO/CMOO concepts in LTE/LTE-A networks is to increase the number of TTIs when all active bearers are satisfied from the viewpoints of different multi-objective criteria. Different scheduling rules work much better when compared with any other disciplines under given circumstances such as channel conditions, queue states, traffic load or QoS requirements. Hence, the principle of applying a mixture of scheduling rules becomes mandatory, and each scheduling discipline is called when it can provide the highest scheduler reward. This chapter summarizes the proposed concepts and results highlighted in the presented research work. The limitations of the proposed scheduling approach are discussed and the possible hardware implementations are proposed in order to show the veracity of the novel scheduling scheme. Some future research directions are presented when the DSR-SMOO/CMOO principles are applied for the heterogeneous traffic and decoupled TDPS/FDPS scheduling.

8.2 Main Results Achieved

The obtained results include the aspects which are focused on the CQI classification techniques and on different types of sequential or concurrent optimization problems. For the CQI aggregation principle, three main stages are

involved: preprocessing, classification and regression. The classification stage includes two steps: unsupervised learning and supervised learning. The unsupervised learning step is an offline procedure in which the best set of pre-processed CQI data centers is obtained. For the supervised learning step, the obtained set of CQI centers is used to train the RBFNN weights. Then, the entire trained structure is exploited in order to classify each CQI report for each active bearer in the corresponding pattern. The aggregate scheduler state space is used to approximate different scheduling rules for given states by using the MLPNN function approximation with different reinforcement functions which define in fact the RL algorithm. Two types of DSR-SMOO problems are analyzed, focusing on the NGMN user fairness and GBR objectives. The proposed DSR-CMOO scheduling policies are focused on the DP and FGDP multi-objective criteria.

8.2.1 LTE Scheduler State Space Aggregation

One of the most important components of the LTE scheduler state space is represented by the subspace of CQI reports. The CQI subspace dimension depends strongly on the number of active users and on the system bandwidth. In order to avoid the bandwidth dependency, a pre-processing technique is proposed in this sense in Appendix C. The CQI preprocessing node aims to reduce the CQI report dimension for each bearer to the number of possible CQI report levels (15 in LTE/LTE-A). Two methods of CQI preprocessing types are proposed in this sense: the *Top Mass CQI* and the *Majority Mass CQI*. The top mass CQI aims to select the best percentages of CQI values for a given system bandwidth whereas the majority mass CQI takes into account the best CQI values which form the majority for a given percentage threshold of the CQI values in the considered bandwidth. The details of these procedures are explained in Appendix B and Appendix C. The idea is to reduce the original CQI state space size to a more comprehensive form which can be collected under a finite number of pre-processed CQI observations for different system bandwidths.

The unsupervised learning step of the CQI state space classification aims to find the most representative preprocessed CQI data centers for a given collection of data points. The idea is to propose a novel k-means clustering

algorithm which is able to minimize the average squared-error distortion between each preprocessed CQI observation and each obtained preprocessed CQI center. The squared-error distortion minimization represents the case of stochastic optimization problem in which the optimal set of CQI data centers is not guaranteed due to the non-convexity characteristic of the clustering constraints. In order to avoid the local minima problems for the determination of the preprocessed CQI data centers, the meta-heuristic concept entitled *Simulated Annealing with Stochastic Tunneling* (SAST) is proposed in this sense. The idea of the proposed method is to find the best set of preprocessed CQI data centers by applying at each stage a mixture of two classical k-means algorithms: Lloyd and Swap heuristics. Based on the temperature factor and tunneling function, SAST takes the advantage of both heuristics by minimizing the mean squared-distortion and by avoiding at the same time, the local minima problem. Simulation results indicated that the SAST meta-heuristic algorithm outperforms other existing approaches such as iterated Lloyd, single Swap, Hybrid-EZ or Hybrid-SA with different numbers of centers. By using the novel Hybrid-SAST meta-heuristic algorithm, the obtained sets of CQI data centers minimize the mean squared-distortion for different CQI collections of different LTE bandwidths. The advantages of using Hybrid-SAST with different numbers of centers for the possible LTE system bandwidths are highlighted in Appendix D and Appendix E.

The supervised learning step aims to train the RBFNN weights based on the obtained set of the preprocessed CQI data centers when the Hybrid-SAST meta-heuristic algorithm is performed. The training stage of the RBFNN weights is conducted through the gradient descent principle by considering two types of observations: training and validation sets. The decision of choosing one of the aforementioned data set at each epoch is taken by using the same SAST meta-heuristic concept in order to minimize the mean-square error between the RBFNN outputs and the considered preprocessed CQI patterns and to avoid the local minima problems which are involved when training the RBFNN weights. In this sense, the optimum set of RBFNN learning rates and Gaussian parameters are determined for a various number of CQI data centers when the system bandwidth is 20MHz. The simulation results have showed that the trained RBFNN structure

is able to minimize the mean squared errors in the testing phase for different combinations of the CQI aggregation schemes, such as $\{Top3, Top4, Top5\}$ with $N_{CT} = \{64, 128, 256, 512, 1024\}$ number of preprocessed CQI centers when the considered system bandwidth is 20MHz. The complete sets of simulation results are presented in Chapter 4 and Appendices B, C and D.

The motivation behind the unsupervised and supervised learning steps is to classify the preprocessed CQI reports in given patterns and to form the classified CQI state space. The classified CQI state size is denoted by the number of centers which is used and contains the number of testing CQI observations belonging to different preprocessed CQI clusters. Based on the obtained classified CQI state space, different statistical methods can be performed as shown in Sub-section 4.6 from Chapter 4. Basically, the overall CQI subspace is reduced to a 4-dimensional CQI vector which can be used by the LTE controller state space.

8.2.2 Sustainable Scheduling Policies Based on the Sequential Multi-Objective Optimization

When the NGMN fairness requirement is considered, the objective of the proposed DSR-SMOO problem is to maximize the mean percentage of TTIs when the scheduler stays feasible and to minimize the mean percentage of TTIs when the scheduler is declared unfair. At the same time, the mean percentage of TTIs with punishment rewards in the exploitation stage should be minimized. The simulation results concluded in this sub-section are presented in Chapter 6, Appendix F and Appendix G.

In the case of the NGMN fairness requirement, the parameterizations of two scheduling techniques of GPF-SP and GPF-DP are performed at each TTI. Eight scheduling policies are obtained when the GPF-SP rule is considered by using different RL algorithms such as Q, DoubleQ, SARSA, QV, QV2, QVMAX, QVMAX2 and ACLA. All these techniques use a predefined set of actions to parameterize the considered optimization problem. CACLA1 uses a single-dimension continuous action in order to find the optimal fairness parameter for the

GPF-SP scheduling rule at each TTI, whereas CACLA2 RL algorithm is performed by using a continuous two-dimensional action space for the GPF-DP parameterization. The AUT observations are calculated by using two types of averaging filters: AUT-EMF and AUT-MMF. When the AUT-EMF observations are considered, the existing methods outperform the proposed policies if the CQI aggregation schemes are not considered. For example, the MT and AS scheduling techniques assure a gain of 3% of feasible TTIs when compared against the best policy which is CACLA2. When the $(Top3, N_{CT} = 64)$ CQI aggregation scheme is exploited, the CACLA2 policy is the best option by outperforming by about 15% the existing proposals in terms of the number of feasible TTIs. When other aggregation schemes are performed, similar results are obtained by minimizing the number of unfair TTIs of about 8% when compared with the MT scheduling technique. The windowing factor plays a crucial role when the DSR-SMOO problems focusing on the NGMN fairness with AUT-MMF observations are considered. For a given domain of windowing factors, CACLA1 and CACLA2 scheduling policies perform the best, indicating a maximum gain of 35% of feasible TTIs when compared with the MT scheme. Also, by using these policies, the percentage of TTIs, when the scheduler is unfair, is minimized by indicating a maximum gain of 31% when compared with the MT existing scheduling scheme when the $(Top3, N_{CT} = 64)$ CQI aggregation scheme is exploited. To conclude, CACLA1 and CACLA2 outperform the existing proposals from the NGMN fairness requirement perspective if and only if the aggregate CQI observations are included in the LTE controller state space. Being able to maximize the mean percentage of feasible TTIs and to minimize the number of punishments in the exploitation stage, the obtained scheduling policies focusing on the NGMN fairness requirement are sustainable.

The DSR-SMOO focusing on the GBR objective considers a mixture of four scheduling rules of GPF-LM, GPF-RAD, GPF-BF and GPF-mM under three traffic types: full buffer, CBR and VBR. When the full buffer traffic type is considered, the proposed scheduling rule GPF-LM assures the best performance for different windowing factor values from the perspective of the GBR satisfied

bearers. The ACLA scheduling policy provides similar results since the obtained policy follows the GPF-LM scheduling rule for the entire downlink transmission. For large windowing factors, ACLA and GPF-LM outperform other scheduling rules by indicating a gain of more than 35% of TTIs when the 100% GBR satisfaction criterion is considered. When the CBR traffic type is analyzed, the best performance is obtained if the ACLA scheduling policy is applied. For large windowing factors, ACLA indicates a gain in percentages of TTIs of about 17% when compared with the main candidates of GPF-LM and GPF-RAD if the 100% GBR satisfaction criterion is taken into account. For the VBR traffic type, ACLA is the best solution for the situation when the active bearers are 100% satisfied from the GBR objective point of view. For lower GBR satisfaction levels, GPF-LM performs better. When large windowing factors are involved in the AUT-MMF computations, GPF-LM shows a gain of about 10% when compared with the ACLA policy in terms of the mean percentage of TTIs with the 80% GBR satisfaction level. The scheduling policies obtained when the ACLA actor-critic learning is performed assure the best performances by maximizing the mean percentage of GBR feasible TTIs and by minimizing at the same time, the mean percentage of TTIs with punishment rewards. The STD values for the obtained results are minimized and then, the proposed scheduling policies are sustainable.

8.2.3 Sustainable Scheduling Policies Based on the Concurrent Multi-Objective Optimization

The concurrent optimization problem focusing on the HoL delay and PDR objectives considers four scheduling rules: GPF-LOG, GPF-EDF, GPF-EXP1 and GPF-EXP2. Due to the particularity of the considered DSR-CMOO problem, the HoL delay constraint is considered to be a percentage from the original HoL delay imposed by the LTE standard. The study aims to analyze the indicated scheduling rule characteristics from the DP multi-objective performance point of view. In this sense, the DP reward function is performed in such a way that the number of TTIs with the maximum rewards differs from the number of feasible states. In other words, when the reward is maximized, the feasible state is not always reached. It

is proceeding in this way in order to verify which scheduling rules aim to reduce the mean HoL delay by minimizing, at the same time, the variation of HoL delays of each active bearer. Simulation results conclude that the GPF-LOG rule indicates the lowest discrepancy between the number of feasible TTIs and the number of TTIs with the highest reward. In this sense, GPF-LOG aims to minimize the mean HoL delay by minimizing, at the same time, the STD indicator of the HoL delays of all active bearers. However, by mixing other scheduling rules with GPF-LOG, the obtained scheduling policies can achieve a notable performance when the DP multi-objective evaluation is considered. The PDR objective performance depends on the selected windowing factor which can set the rate of dropped packets to be matched in a given constraint for a shorter or larger time window. In the case of the CBR and VBR traffic types, the simulation results indicate that by mixing the considered rules to be applied on the best matching conditions, the obtained scheduling policies are able to outperform the existing scheduling techniques from the viewpoint of 100% DP satisfied bearers when different settings for the windowing factors are considered with a minimum variation of the obtained results. The amount of punishment rewards is minimized revealing in this way the sustainability of the obtained scheduling policies.

When the DSR-CMOO problems consider the performance of the NGMN fairness, GBR, PDR and the HoL delay objectives, the obtained scheduling policies combine the entire set of analyzed scheduling rules. The system architecture suffers a slight modification since two types of controllers are considered: the fairness controller and the QoS controller. The type of multi-agent based architecture considers the model of cooperation only when the GPF-DP scheduling rule is selected by the QoS controller as indicated in Chapter 5. In this case, the CACLA2+ RL approach adapts the fairness parameters and the windowing factor at each time when the GPF-DP is selected and the fairness controller is implicitly called in this sense. In the reward function computation, two sets of parameters are considered to be very important: the particular reward weights and the intrinsic reward decision parameters. The weights of the QoS particular rewards are very important since these parameters select the importance of each objective in the multi-objective reward function. For this study, it is

assumed that each particular reward has the same weight and implicitly requires an equivalent priority in the global reward computation. Future research directions may include different settings of these parameters for different traffic types. The second type of parameters refers to the temporal difference between two intrinsic reward selections when the global reward function is computed. In this study, only the global reward computation considers at each TTI the difference between two consecutive intrinsic rewards.

When the CBR traffic type is considered during the exploitation stage, the proposed set of scheduling policies outperforms the existing approaches for different performance domains of the FGDP criterion. The SARSA policy is the best choice when the percentage of $\overline{p}_{TTI}^{FGDP,100\%}$ is analyzed, whereas the QV2 policy performs better for the $\left\{\overline{p}_{TTI}^{FGDP,80\%}, \overline{p}_{TTI}^{FGDP,85\%}\right\}$ levels. The mean percentage of TTIs with the FGDP punishment rewards for SARSA policy indicates a level of $\overline{p}_{TTI}^{FGDP,MRW} = 26.44\%$ which represents the best performance when compared against other RL policies. For the VBR traffic type, the SARSA policy gains more than 5% of $\overline{p}_{TTI}^{FGDP,100\%}$ when compared with other existing techniques whereas Q-L and QV2 achieve the highest amount of TTIs when the $\left\{\overline{p}_{TTI}^{FGDP,80\%}, \overline{p}_{TTI}^{FGDP,85\%}\right\}$ performance metrics are considered. The SARSA scheduling policy for the VBR traffic type obtains a mean percentage of TTIs with FGDP punishment rewards of about $\overline{p}_{TTI}^{FGDP,MRW} = 44.34\%$ which is considered the best performance among the learned RL policies. For all considered scenarios, the simplest CQI aggregation architecture is exploited in terms of the $\left(Top3, N_{CT} = 64\right)$ aggregation scheme by using the optimized RBFNN parameters obtained in Chapter 4.

The simulation results are conducted through fluctuating traffic load, QoS requirements and under most severe channel conditions in terms of Jakes fast fading models. Under these circumstances, the obtained scheduling policies maximize the mean percentage of FGDP feasible TTIs while minimizing the mean percentage of TTIs with FGDP punishment rewards. The STD values for

both types of performance indicators are minimized. Therefore, the obtained scheduling policies being focused on the FGDP multi-objective criterion are sustainable. The extended sets of simulation results for the DSR-CMOO problems are presented in Chapter 7 and Appendix H.

8.2.4 Publications Arising From This Work

Journal Papers

I.-S. Comşa, M. Aydin, S. Zhang, P. Kuonen and J. F. Wagen, "Multi Objective Resource Scheduling in LTE Networks Using Reinforcement Learning," in *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 3(2), pp. 39-57, April 2012.

Conference Papers

I.-S. Comşa, S. Zhang, M. Aydin, J. Chen, P. Kuonen and J. F. Wagen, "Adaptive Proportional Fair Parameterization Based LTE Scheduling Using Continuous Actor-Critic Reinforcement Learning," *in IEEE Global Communications Conference (GLOBECOM)*, pp. 4387 - 4393, Dec. 2014.

I.-S. Comşa, M. Aydin, S. Zhang, P. Kuonen, J. F. Wagen and L. Yao, "Scheduling Policies Based on Dynamic Throughput and Fairness Tradeoff Control in LTE-A Networks," in *39th Annual IEEE Conference on Local Computer Networks (LCN)*, pp. 418-421, Sept. 2014.

I.-S. Comşa, S. Zhang, M. Aydin, P. Kuonen, and J. F. Wagen, "A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-Advanced Technologies Using Neural Networks, " in *37th Annual IEEE Conference on Local Computer Networks (LCN)*, pp. 332-335, Oct. 2012.

I.-S. Comşa, M. Aydin, S. Zhang, P. Kuonen and J. F. Wagen, "Reinforcement Learning Based Radio Resource Scheduling in LTE-Advanced," in *17th International Conference on Automation and Computing (ICAC)*, pp. 219 – 224, Sept. 2011.

8.3 Limitations of the Proposed Approach

The main issue when proposing the set of sustainable policies for the DSR-SMOO/CMOO problems refers to the trade-off between exploration and exploitation when the RL approach is used. The set of results provided in this study is based on multiple simulations and then, the optimum number of TTIs in the exploration and exploitation stages is determined. Consequently, other parameters of the MLPNN structures such as the number of hidden layers and the number of hidden nodes are determined based on the number of observations provided in the exploration stage. Unfortunately, there is no way in LTE scheduling to determine the exact number of TTIs for the exploration stage and when exactly the exploitation stage should start. The only method which can help to overcome this aspect is to provide the collection and preprocessing stages before the exploration stage. In the preprocessing stage, the controller states should be collected and only the most representative observations with the best characteristics should be stored. Then, the exploration stage is an offline procedure which trains the MLPNN weights based on the set of collected controller states. When the scheduling policies are refined and improved enough based on the finite controller state space, then the exploitation stage can be performed by using the observations obtained from the real-time LTE network.

8.4 Possible Hardware Architectures

The implementation of the proposed approaches refers to the integration of the trained RBFNN and MLPNN functions on the real hardware architectures. The LTE-A Scheduler can be used to train the RBFNN non-linear functions for the CQI state space aggregation as presented in Chapter 4. Once the set of the preprocessed CQI data centers for each bandwidth and the set of RBFNN weights are trained and fixed, then the RBFNN classification structure can be implemented by using the Field Programmable Gate Array (FPGA) or the Very High Speed Integrated Circuit Hardware Description Language (VHDL) as indicated in [247], [248], [249], [250]. The advantage of using such architectures

is to exploit the parallelism of these circuits in order classify each CQI report for each active user in parallel based on the sets of preprocessed CQI centers and based on the RBFNN weights provided from the LTE-A Scheduler simulator.

When the RL approaches are used to train the MLPNN non-linear functions in order to approximate the scheduling rules for different scheduler states, the LTE-A Scheduler provides the sustainable set of scheduling policies as shown in Chapter 5, Chapter 6 and Chapter 7. Once the MLPNN weights are fixed and implicitly the scheduling policies are refined enough, then the same FPGA and VHDL hardware architectures can be used to implement the learned MLPNN non-linear functions for each scheduling rule. Based on these approaches, the parallelism can be exploited in order to feed-forward the aggregate scheduler state in parallel for each MLPNN function and for each scheduling rule at each TTI.

8.5 Future Directions

One possible research direction is to include the present work in the presence of multiple traffic types. Basically, there are two ways of treating this innovative approach: to use the existing policies or to train other scheduling policies for each traffic type. The ideas to be presented in Sub-section 8.5.1 aim to train different scheduling policies based on different traffic types by using slave controllers whereas the scheduling decision on the traffic classes is performed by the master controller. The second proposal refers to the possibility of integrating the RL principle in the decoupled TDPS/FDPS architectures as discussed in Chapter 2. Sub-section 8.5.2 proposes the future research directions in this sense.

8.5.1 RL for the DSR-SMOO/CMOO Scheduling with Traffic Priorities

For more realistic LTE scheduling scenarios, the scheduling procedure is performed in the presence of multiple traffic types (i.e., heterogeneous traffic). The traffic heterogeneity implies different QoS requirements (as shown in Table 2.1 from Chapter 2). In general, the reward functions for different objectives such

as $\mathcal{RW}_t^D$ or $\mathcal{RW}_t^G$ represent the sum of normalized sub-rewards from different traffic classes. Therefore, the traffic heterogeneity requires different reward functions for different priority classes.

The controller state space dimension becomes much larger than the classical state space with homogeneous traffic. Let us define $\mathcal{S}_t^{C,p}$ as the state space for traffic class p, $p = 1,..,N_P$, where N_P is the number of traffic classes or priorities. Then, the overall state space dimension, when the DSR-SMOO/CMOO MDP with heterogeneous traffic is considered, becomes:

$$\mathcal{S}_t^{C,P} = \bigcup_{p=1}^{N_P} \mathcal{S}_t^{C,p}, \;\; D\left[\mathcal{S}_t^{C}\right] = N_P \cdot D\left[\mathcal{S}_t^{C,p}\right], \;\; p = 1,...,N_P \tag{8.1}$$

As mentioned earlier, the priority based total reward function can be defined as a weighted sum of the total reward for each class as indicated in Equation 8.2:

$$\mathcal{RW}_t^{T,P} = \sum_{p=1}^{N_P} w_p \cdot \mathcal{RW}_t^{T,p} \tag{8.2}$$

where w_p indicates the importance of each reward based on the priority table being defined by the LTE standard. More details about the multi-objective rewards $\mathcal{RW}_t^{T,p}$ from the perspective of different objective combinations are provided in Chapters 6 and 7.

The action and the state spaces of each slave agent are similar to the case of DSR-CMOO combinatorial problems with the homogeneous traffic type. The main problem is the state space dimensionality of the master agent which requires a high computational complexity for the MLPNN function approximations and expensive exploration time. The current approach aims to avoid these drawbacks by partitioning the original state space in N_P subspaces. Each subspace represents the slave agent state space for a given traffic class. The slave agents can be trained separately from the master agent. When the QoS and fairness agents for each slave controller are trained enough, then these structures can be exploited when the master agent is trained. The state space portioning is achieved based on the

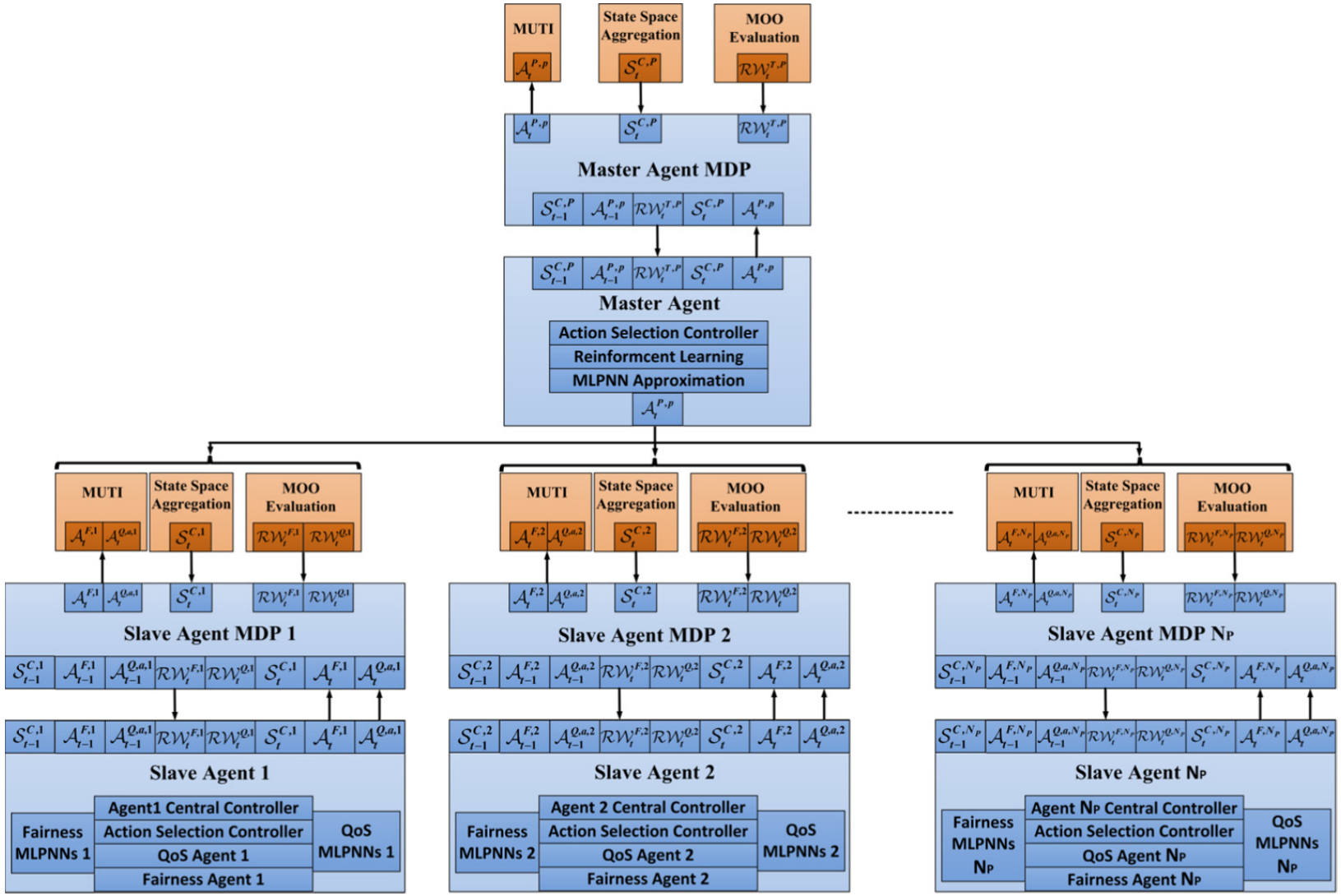


Fig. 8.1 The RL Based DSR-SMOO/CMOO MDP with Heterogeneous Traffic

discrete action set $\mathcal{A}_t^P \in \mathbb{N}$ which represents the traffic class index (1-highest priority, N_P-lowest priority). In fact, $\mathcal{RW}_t^{T,P}$ is the quality measure of applying action $\mathcal{A}_{t-1}^p$ at TTI *t*-1. The MDP problem for the master controller becomes:

$$\left[\left\langle \mathcal{S}_{t-1}^{C,p}\right\rangle,\left\langle \mathcal{A}_{t-1}^{p}\right\rangle,\left\langle \mathcal{A}_{t-1}^{F,p},\mathcal{A}_{t-1}^{Q,a,p}\right\rangle,\mathcal{RW}_t^{T,P},\left\langle \mathcal{S}_{t}^{C,p}\right\rangle,\left\langle \mathcal{A}_{t}^{p}\right\rangle,\left\langle \mathcal{A}_{t}^{F,p},\mathcal{A}_{t}^{Q,a,p}\right\rangle\right]_{\substack{a=1,..,\left|\mathcal{A}^Q\right| \\ p=1,..,N_P}} \tag{8.3}$$

The idea of the heterogeneous state space partitioning is to divide the MDP problem from Eq. 8.3 into N_P MDP sub-problems which correspond to different priority classes such that:

$$\left[\mathcal{S}_{t-1}^{C,p},\mathcal{A}_{t-1}^{F,p},\mathcal{A}_{t-1}^{Q,a,p},\mathcal{RW}_t^{T,p},\mathcal{S}_{t-1}^{C,p},\mathcal{A}_{t}^{F,p},\mathcal{A}_{t}^{Q,a,p}\right]_{\substack{a=1,..,\left|\mathcal{A}^Q\right| \\ \forall p=1,..,N_P}} \tag{8.4}$$

The obtained MDP sub-problem in Eq. 8.4 for the traffic class *p* can be solved by using the methodology proposed in Chapter 5. At each TTI, the scheduler controller performs two main steps:

1. **Senses** the overall state space $\mathcal{S}_t^{C,P}$ and select a subspace $\mathcal{S}_t^{C,p}$ based on the selected priority action $\mathcal{A}_t^p$. This level of the general controller architecture is entitled ***Master Agent***.
2. Based on the master agent decision, the selected subspace is passed to the second level of the scheduler controller. The second level has to solve the MDP problems defined by Eq. 8.4. The entity which operates at this level is entitled ***Slave Agent***. The scheduler controller has N_P slave agents. At each TTI one or multiple slave agents can be selected (and the scheduling procedure considers multiple bearers with different priority levels).

In fact, each slave agent is the MARL architecture with specific cooperation between two sub-agents: QoS and fairness agents. The slave agents are rewarded based on the MOO performance for each class, whereas the master agent is rewarded by using the sum of weighted rewards from Equation 8.2.

On the scheduler side, the MUTI entity has to perform the scheduling procedure by selecting the traffic class indicated by $\mathcal{A}_t^p$ in order to parameterize

the fairness MU based on $\mathcal{A}_t^{F,p}$ and to select a proper scheduling rule based on $\mathcal{A}_t^{Q,a,p}$. Of course, if there is not enough data to transmit in one selected class for the whole bandwidth, the scheduler can decide to schedule the next priority class after allocating the best RBs to the selected class. The diagram block for DSR-SMOO/CMOO MDP problems with the heterogeneous traffic is highlighted in Fig. 8.1. The slave controllers are not updated at each TTI. The updating period depends on the time when, for example slave agent *p* was scheduled last time.

One way to reduce the computational complexity for the hierarchical structure exposed in Fig. 8.1 is to train first the slave agents for very dynamic conditions (traffic load, QoS requirements) and then, to use the trained structure for the second phase when the MLPNN functions of the master agent are trained.

8.5.2 RL in Decoupled TDPS/FDPS Scheduling

The system architecture from Chapter 2 and the multi-objective optimization model from Chapter 3 are proposed for the coupled TDSP/FDPS scheduling, where the active user selection and the resource allocation are jointly achieved by using one scheduling rule which is decided at each TTI by different RL algorithms in the exploration or exploitation stages. It is important to remind that in the decoupled TDPS/FDPS scheduling, the user selection is performed first in time domain based on the particular TDPS scheduling rule (QoS scheduling rule), and then, the selected group of users is passed in the frequency domain in order to be scheduled by using the FDPS scheduling rule (GPF-SP/DP).

One possible research direction is to propose sets of sustainable scheduling rules in order to address the multi-objective problem separately: the GBR, HoL delay, PDR and queue stability in the TDPS domain and the user fairness and system throughput tradeoff in the FDPS domain. The mathematical model exposed for the coupled TDSP/FDPS-DSR in Chapter 3 is simplified since the objective and the scheduling rule are determined in the TDPS domain and the FDPS domain addresses only the simple RB allocation optimization problem under variable fairness parameters (α_t, β_t) which have to be adjusted at each TTI.

When the TDPS-SSR/FDPS-DSR scheduling model is used, the active users are selected in the time domain based on the static scheduling rule focusing on one or multiple objectives such as: GBR, HoL delay, PDR or queue stability. In the frequency domain, the DSR-SMOO problem from Chapter 6 is solved by using the same principles of CACLA1 or CACLA2 to parameterize the fairness variables. Both types of observations AUT-EMF and AUT-MMF can be used in order to analyze the performance of the obtained scheduling policies. The state space and the reward function remain similar to those proposed in Sub-section 6.2.3 from Chapter 6 with the amendment that, the system complexity is much lower when higher traffic load is scheduled. The TDPS-DSR/FDPS-SSR scheduling implies the fact that the multi-objective optimization can be achieved in the time domain by performing different TDPS rules which are focusing on different QoS targets. The controller state elements keep similar to those proposed in Sub-section 7.3.3 and the reward function can be computed similarly to Equation 7.23 with a slight difference in the sense that, the reward functions for the system throughput and user fairness are not included. In the FDPS domain, a static parameterization scheme is applied in terms of the GPF scheduling rule.

Two types of controllers are required when the TDPS-DSR/FDPS-DSR scheduling scheme is used: the QoS controller selects the proper scheduling rule based on the RL algorithms with discrete action spaces in order to select different users based on their QoS budget. The fairness controller performs CACLA1 or CACLA2 RL approaches in order to stabilize the normalized throughput observations under the NGMN fairness criterion. There is no specific cooperation between the agents and different controller state spaces are used by the QoS and fairness agents $\left(\mathcal{S}_t^{C,F} \not\subset \mathcal{S}_t^{C,GDP}\right)$. Then, one reward function is received by the QoS agent and one reward function from Eq. 6.17 and Eq. 6.18 is received by the fairness agent. In this sense, the fairness agent is updated at each TTI. The traffic prioritization can be solved in the time domain scheduling where different users with different QoS profiles can be selected. The main drawback of these techniques refers to the fact that the proposed architectures are suboptimal and in general, the system throughput is seriously degraded when compared with the coupled TDPS-FDPS-DSR scheduling architectures.

Appendix A

Related Studies on the MOO-Based LTE Scheduling

A.1 Appendix Outline

In this section, the related studies on the SSR-SMOO/CMOO scheduling problems are discussed in order to highlight the necessity of the proposed scheduling schemes in order to overcome the drawbacks of the existing methodologies. In Section 3.8 from Chapter 3, the main related studies concerning the SSR-SMOO/CMOO problems in the coupled TDSP/FDPS scheduling are presented based on the proposed classification scheme from Section 3.7. Then, in this section, the extended related studies are discussed in terms of the SSR-SMOO/CMOO problems in the coupled and decoupled TDPS/FDPS scheduling. The rest of the appendix is organized as follows: Section A.2 presents the related work on the SMOO problems focusing on the system throughput, Section A.3 presents the SMOO methodologies being oriented on different fairness criteria, Section A.4 introduces the existing works concerning the SMOO problems focusing on the GBR requirement, Section A.5 presents the SMOO scheduling being focused on the HoL packet delay and Section A.6 highlights the related work of the SMOO problems focusing on the queue stability. In Section A.7, the

existing work for the CMOO problems focusing on the multi-objective criteria is presented and finally, Section A.8 summarizes the state of the art in LTE scheduling based on the classification scheme proposed in Section 3.7.

A.2 SMOO Focusing on the System Throughput

The LTE scheduler being focused on the system throughput maximization should solve the sum of rates optimization problem in an optimal manner. By using the typical linear integer programming models, the complexity cost is very high. This is the case of MCS and RBs assignments for the MT scheduling rule. The authors divide the non-linear optimization problem into two sub-optimal linear sub-problems in [64], where the RBs assignation is performed in the first instance, and then, for the allocated RBs, different MCS schemes are assigned. The original problem has a complexity of $\mathcal{C}\left(|\mathcal{U}_t| \times |\mathcal{B}| \times N_{MCS}\right)$, where N_{MCS} is the number of MCS schemes. The degradation in system throughput increases with the number of users, but the scheduler complexity is considerably reduced to $\mathcal{C}\left(|\mathcal{U}_t| \times |\mathcal{B}| + N_{RB}^{UE_i} \times N_{MCS}\right)$, where $N_{RB}^{UE_i}$ represents the number of RBs allocated to UE $i \in \mathcal{U}_t$ at each TTI. The simulated annealing is proposed in [73] in order to solve the initial linear sum-rate optimization problem. The main idea is to find the most suitable values for: a) the decision vector for users and b) the decision matrix for MCS and RB assignments. The idea of the simulating annealing principle is to start from the original solution of the decision vectors and to search for different temperature levels new solutions in such a way that the optimization problem is maximized. The solution is iterated at each temperature level for a number of times based on the proposed neighborhood function. When the function is performed, a new solution for the decision vectors is generated and the process is repeated until the best solution is reached. Important is the fact that the system complexity is reduced when compared with original problem. The results shows that SA approach is able to provide near-optimal solution, but still with high computational overhead. However, the same authors proposed a Genetic Algorithm (GA) in order to solve the same initial linear integer programming

problem [74]. The idea is to have different solutions in terms of the scheduling decision vector and decision matrix, and to recombine the solutions from different initial couples and to generate more useful solutions called child solutions or chromosomes. The mutation operator is used to evolve the population of each child solution and the crossover operator is used to recombine different child solutions from at least two parent solutions. A chromosome is a particular solution in terms of the user decision vector and the MCS and RBs matrix decision. The mutation operator is based on the neighborhood function used by the SA approach. The crossover operator creates new chromosomes of new child solutions by copying the complete information of odd-numbered columns user-by-user and not RB-by-RB. More precise, for each user index the entire column containing the RBs is copied from one child to another but at different user index. In this way the system complexity is reduced to $\mathcal{C}\left(\left|\mathcal{U}_t\right|\right)$. From the viewpoints of the average bit rate and system complexity, the GA approach provides the best alternative when compared with SA and sub-optimal solutions. The problem exposed above is a *static* and *deterministic optimization problem*.

The problem of coupled TDPS/FDPS scheduling subject of instantaneous sum-power constraint is analyzed in [75]. The objective is to maximize the sum-rate optimization problem $\left(P_T\right)$ from Equation 3.31 from Chapter 3, in which the ACK/NACK feedbacks are considered as component of the scheduler state space. The authors prove that the optimal solution can be found by using the Partially Markov Decision Process (POMDP) which is very hard to be implemented in practice. Then, an upper bound of POMDP problems known as *causal global genie* has been proposed. A greedy coupled TDPS/FDPS approach is developed in order to keep a posterior distribution for each RB with a polynomial complexity. The particle filtering method is proposed to update the posterior distribution at each TTI based on received ACK/NACK feedbacks. The results show a near-optimal solution by providing at the same time an improvement in the system capacity when compared with the uniform power allocation optimization problem, but, with the price of higher system complexity.

Other important research direction is the opportunistic LTE scheduling under imperfect or limited CQI feedback information. This problem is investigated in [76], in which the estimation of the channel information is achieved based on the exploitation of memory inherent Markov channels correlated with the ACK/NACK feedbacks. The non-linear optimization problem $\left(P_T\right)$ is divided into two sub-optimal problems: 1) the channel estimation and rate adaptation in order to maximize the expected instantaneous scheduled user rate and 2) the coupled TDPS/FDPS scheduling based on the optimized achievable user rates. The scheduling problem is modeled as a POMDP problem in which the exploration and exploitation stages are difficult to be defined in order to reach sum-throughput near optimal solutions. In this sense, the Restless Multi-armed Bandit Processes (RMBP) with the Whittle's policy [77] is proposed. Similar to the RL methodology, the Whitlle's analysis considers states, actions and rewards. The state is represented by the average user rate, channel state and the estimator and rate adapter pair. Then the immediate reward function is computed as a performance measure of channel estimation and rate adaptation. The action decides if the user should be scheduled or not. Therefore, the optimal scheduling policy maximizes the infinite horizon discounted reward which is similar to the aggregate problem optimization $\left(P_{Dual2}^{'Agg1}\right)$ from Equation 3.62, Chapter 3. The proposed work follows the greedy policy due to the fact that the approach schedules that user with the highest belief value. The belief value is given by the Bellman equation [78] similar to the updating equations of the state and state-action values (more details are presented in Chapter 5). By exploiting the channel memory, the Whittle's index analysis provides a near-optimal solution of the channel estimation and the scheduling optimization problem.

A.3 SMOO Focusing on User Fairness

Another interesting aspect is the study of the fairness performance under sub-optimal decoupled TDPS/FDPS architectures. In [92] different static scheduling rules are compared for different sub-optimal stages. In this sense, a

new scheduling rule for the FDPS stage is proposed here, belonging to the first class of utilities as shown by Eq. A.1.

$$\begin{cases} U_{2,i}^{1}\left(r_i[t]\right)=\left(r_i[t]\right)^2 \Big/ \left(2\cdot \hat{r}_i[t]\right) \\ F_{2,i}^{'1}\left(r_i[t]\right)=r_i[t] \\ W_{2,i}^{1}\left(\hat{r}_i[t]\right)=1/\hat{r}_i[t] \\ \hat{r}_i[t]=\sum_{j=1}^{|\mathcal{B}|} r_{i,j}[t] \\ D_{2,i}^{1}\left(\hat{r}_i[t]\right)=r_{i,j}[t]/\hat{r}_i[t] \end{cases} \tag{A.1}$$

The obtained scheduling rule in Eq. A.1 is entitled Throughput to Average (TTA). TTA addresses the fairness performance and it is expected to enhance the fairness performance when compared with the simple PF rule since the short term fairness performance is addressed. However, authors in [92] conclude that TDPS-MaxFair/FDPS-TTA outperforms other schemes such as TDPS-PF/FDPS-PF and TDPS-MT/ FDPS-MT from the fairness performance point of view at the expense of system throughput degradation. The same principle of the decoupled TDPS/FDPS is studied in [93]. The difference is that the proposed scheduler supports mixed traffic of BE and CBR under the GBR requirements. The TDPS scheduling divides users into two sets: the first set has the highest priority and contains users with data rates below their GBR targets, and the second set represents users with fulfilled GBR requirements. At the beginning of each TTI, the TDPS scheduler decides which set should be scheduled and then, based on the selected set, different rules are applied: MaxFair for the first set and PF in time domain for the second one. For these reasons, this stage is entitled the DSR based Priority Set Scheduler based TDPS (TDPS-DSR-PSS). For the FDPS scheduler, different static rules are analyzed and compared, such as PF, TTA and PFSch. The

$$\begin{cases} U_{4,i}^{2}\left(\overline{T}_i^{Sch}[t]\right)=U_{2(\alpha,\beta),i}^{2}\left(\overline{T}_i^{Sch}[t]\right) \\ F_{4,i}^{'2}\left(\overline{T}_i^{Sch}[t]\right)=1\Big/\left(\overline{T}_i^{Sch}[t]\right)^{\alpha} \\ W_{4,i}^{2}\left(r_{i,j}[t]\right)=\left(r_{i,j}[t]\right)^{\beta-1} \\ D_{4,i}^{2}\left(\overline{T}_i^{Sch}[t]\right)=\left(r_{i,j}[t]\right)^{\beta}\Big/\left(\overline{T}_i^{Sch}[t]\right)^{\alpha} \end{cases} \tag{A.2}$$

PFSch $(\alpha = 1, \beta = 1)$ scheduling rule represents a special case of the GPF discipline and follows the form expressed in Eq. A.2 known as GPFSch-DP, where $\overline{T}_i^{Sch}[t]$ is updated only and only if the UE *i* has been scheduled in the previous TTI [94]. The same parameter is used for the GPF-RAD scheduling rule which is introduced in Eq. 3.69, in Chapter 3. The updating formula is illustrated in Equation A.3 such as:

$$\overline{T}_i^{Sch}[t] = \begin{cases} \left(1-\beta_{\overline{T}^{Sch}}\right)\cdot \overline{T}_i^{Sch}[t-1] + \beta_{\overline{T}^{Sch}} \cdot T_i[t], if \;\; UE_i(t-1) \\ \overline{T}_i^{Sch}[t-1], if \;\; UE_i\left(\overline{t-1}\right) \end{cases} \quad \text{(A.3)}$$

where the scheduled average user throughput $\overline{T}_i^{Sch}[t]$ can be considered as a part of controllable scheduler subspace $\mathcal{S}_t^{S,C}$ and $UE_i(t-1)$ suggests the fact that UE *i* has been scheduled at TTI *t*-1 and $UE_i\left(\overline{t-1}\right)$ for the opposite case. The TDPS-DSR-PSS/FDPS-PFSch scheduling method enhances the coverage performance of about 60% when compared with the TDPS-DSR-PSS/FDPS-PF scheme at the price of a cell throughput loss of about 5% [93]. The TDPS-DSR-PSS/FDPS-SSR scheme represents the case of CMOO optimization, but it is related in this section because the presented results are focused only on the user fairness and system throughput tradeoff.

A.4 SMOO Focusing on the GBR Objective

In [96] a low complexity heuristic algorithm is used to solve the initial non-linear mixed-integer optimization problem subject to rates constraints. The algorithm embraces, at each TTI, two phases: the first one is based on prediction, and for each user, the CQIs are sorted in the descending order and the execution is performed at each TWRG; in the second one, based on the sorted list of each UE, at each TTI a function value is calculated to determine the number of transmission opportunities that each UE would need until the end of TWRG in order to meet its data rate requirement. The number of RBs that can be allocated to each user is calculated at each TTI. The RB assignment is achieved based on the TTA

scheduling rule constrained by the maximum number of allowable RBs. The proposed approach performs very close to the optimal solution, achieving a better performance in guaranteeing the user rate requirements when compared with classical approaches such as PF, Weighted RR or MT static rules.

The most adopted scheduling policies translate the non-linear integer optimization problem into linear integer optimization problem by mapping the QoS constraints into specific and optimal utility functions as discussed in Section 3.5 from Chapter 3. Based on the above principle, the rate guarantee can be addressed in terms of the residual time. The residual time t_i^{RES} is an urgency measure that defines the time in which one flow can wait in the queue without violating the GBR or other QoS requirements [97]. The residual time t_i^{RES} parameter is determined based on the queue length and based on the transmission history for each flow and it can be included in the list of scheduling rules being focused on the GBR requirement such as:

$$\begin{cases} U_{5,i}^{3}\left(\overline{T}_i[t]\right) = \left(1/t_i^{RES}[t]\right)\cdot U_{2(\alpha,\beta),i}^{2}\left(\overline{T}_i[t]\right) \\ F_{5,i}^{'3}\left(\overline{T}_i[t]\right) = 1/\left(\overline{T}_i[t]\right)^{\alpha} \\ W_{5,i}^{3}\left(t_i^{RES}[t]\right) = \left(1/t_i^{RES}[t]\right)\cdot\left(r_{i,j}[t]\right)^{\beta-1} \\ D_{5,i}^{3}\left(t_i^{RES}[t]\right) = \left(1/t_i^{RES}[t]\right)\cdot\left(r_{i,j}[t]\right)^{\beta}/\left(\overline{T}_i[t]\right)^{\alpha} \end{cases} \quad \text{(A.4)}$$

By using the method of Lagrange multipliers in the original mixed-integer programing optimization problem [47], [102], [105], the rate constraints introduce dual variables in the optimization problem, and the obtained scheduling rule is entitled Stochastic Primal Dual (SPD). If the GPF is used as the utility function, then the scheduling scheme becomes GPF-SPD subject to rate constraints [47]. An innovative scheduling rule is introduced in Chapter 6 being entitled the GPF-LM discipline. This rule is able to provide the best results in terms of the mean percentage of feasible TTIs for the GBR objective when the full buffer traffic type is simulated. Under the CBR and VBR traffic types, the GPF-LM rule shows its limits in the detriment of other obtained scheduling policies. More details about these concepts are provided in Chapter 6.

Throughput guarantee adaptive scheduling schemes based on the violation probability have been proposed in [103]. The violation probability refers here to the probability of not fulfilling the requested bit rate in a given TWRG. The simulation results indicate that the proposed scheme performs better when compared with MT, PF, RR scheduling rules in terms of the throughput guarantee for homogeneous and heterogeneous GBR constraints. The same schemes are studied in [104] for imperfect CQI reports. By using an optimal maximum a posteriori predictor for the noisy and the delayed CQI reports, the authors conclude that the outage probability cannot be reduced to zero for the proposed schemes when the imperfect CQI reports are considered.

A.5 SMOO Focusing on HoL Delay Objective

Two-level downlink scheduling with different time granularities is proposed in [117], [118]. The TDPS is performed at each frame (10ms), based on the discrete time control theory, the amount of RT data to be transmitted in the next frame is calculated in order to satisfy the delay constraints. The MT rule is used by the FDPS domain for the RT scheduling and PF for the BE flows. The frame level TDPS (FL-TDPS) performs better than GPF-EXP2 and GPF-LOG when $(\alpha = 1, \beta = 1)$ from the PLR point of view [117], [118].

A decoupled TDPS/FDPS scheme focusing on HoL requirements for different classes of services is provided in [119]. In the TDPS state, an inter-class resource distribution is performed based on the cooperative game coalition between different classes with the sigmoid function utility (U-delay) that performs the resource distribution. The Lagrange multiplier is used to provide the Pareto Optimality [119]. In the FDPS state, the scheduling is achieved thorough delay prioritized scheduling (DPS) [120]. Users with the minimum DPS metric are selected for the transmission and the RBs with the highest SINRs are selected for transmission. Users that approach to the HoL deadline receive the best channel conditions for the transmission. From the system delay point of view, U-delay performs much better than GPF-MLWDF with $(\alpha = 1, \beta = 1)$ in some conditions

for mixed video, CBR, VoIP and gamming traffic types. On the other hand, U-delay outperforms GPF-MLWDF, GPF-EXP2 and GPF-DP rules when the fairness parameters are $(\alpha = 1, \beta = 1)$ from the viewpoints of the system throughput for each traffic class, PLR and user fairness even if the main target is the HoL packet delay satisfaction.

A very important aspect of the decoupled TDPS/FDPS schemes is the scheduling procedure of VoIP users in the presence of other traffic types such BE. The semi-persistent scheduling aims to allocate different resources for VoIP users. The priority mode controls the duration of semi-persistent scheduling in order to increase the system throughput and to avoid the starvation of other traffic classes [121]. By using the priority mode deals with the low resource utilization due to the small sizes of VoIP packets or due to the fact that the allocated resources may be not sufficient due the poor channel conditions. Then, a coupling method permits two VoIP users to share the resources allocated to them [122]. In [123] is proposed a novel priority mode coupling method in the TDPS stage in which pairs of users with opposite channel conditions are allowed to share their resources. The FDPS stage is performed by using the Channel Adaptive Fair Queuing (CAFQ) [124]. The simulation results show the benefit of the proposed method in comparison with the existing approaches by minimizing the PDR for the VoIP users. The semi-persistent scheduling schemes can be applied for the proposed RL methodology by using the principle of the passive user selection and active resource allocation which is exposed in Sub-section 2.9.1 from Chapter 2. Other aspects of semi-persistent scheduling can be found in [125], [126].

In [127] is proposed an intelligent decoupled TDPS/FDPS scheduling scheme focusing on the PLR constraint. In TDPS, the Hebbian learning [128] is used to distribute the RBs among the real time and non-real time traffic types while the k-means clustering algorithm sorts and prioritizes different groups of real time users based on their PLR performance. The group of real time data flows with the lowest distance from the centroid to the PLR requirement vector is preferred to be scheduled in the FDPS domain. The FDPS domain can use simple scheduling rules such as MT, PF or MaxFair. The proposed method is able to

reduce the average PLR by guaranteeing, at the same time, the system throughput improvement for the non-real time traffic types.

A.6 SMOO Focusing on the Queue Stability

One of the most problematic issues in the QoS guaranteeing scheduling is the queue stability. The queues need to be stabilized due to a various stochastic processes such as fading radio channels and arrival rates models. Basically, the arrival packet rate should not be greater than the scheduling rate. More precise, authors in [130] conclude that if the average arrival rate lies within the capacity region, then the queue is considered to be stable. This condition implies that $\overline{T}_i[t] \geq \overline{\lambda}_i[t], \forall i \in \mathcal{U}_t$ represents a necessary and a sufficient condition to keep the queue stable for each user $i \in \mathcal{U}_t$. The opposite condition represents the congestion case, and the stabilization process implies the mechanism of moving the mean arrival rates within the achievable rate region.

The congestion control problem becomes a crucial task under the flow-level dynamic when the traffic load fluctuates dramatically during the scheduling period. The MaxWeight rule [131] is considered to be throughput optimal when the number of flows remains constant within the whole transmission. However, when the traffic load changes, MaxWeight is not able to assure the throughput optimality, driving at the same time, the system into the instability zone [132]. In [131] is proposed a new scheduling scheme called workload-based scheduling with learning (WSL) based on Foster-Lyapunov drift which does not need to know the prior knowledge of the channel conditions and the traffic load. The results show that the congestion probability is much lower under the WSL scheme when compared with the MaxWeight scheduling.

Other approach integrates the frugality constraint from the optimization problem into the scheduling rule [49], [50]. The rule is known as GPF-MDU with $(\alpha = 1, \beta = 1)$ and it is presented in Section 3.5.5 from Chapter 3. In [49] the authors defined the Maximum Stability Region (MSR) which can be achieved by one scheduling policy under all other existing policies. If the MUF is polynomial,

then it is sufficient for the GPF-MDU to guarantee the MSR region as indicated in [49], [50]. Under various mixed traffic various loads, GPF-MDU outperforms other rules such GPF-MLWDF, GPF-PF and GPF-EXP2 rules when $(\alpha = 1, \beta = 1)$ from the mean system delay point of view, assuring at the same time the system stability [49], [133], [134].

The problem of stability in the presence of elastic and RT traffic types is studied in [135]. It is shown that the MSR reduces significantly when the number of best effort flows increases. The reduction of the stability region is proportional to the opportunistic gains (when MT is used), and the admission control for best effort users is suggested in order to avoid the local instability [135].

A.7 CMOO Focusing on Multiple QoS Objectives

As indicated in Fig. 3.3 in Chapter 3, there are two main ways of addressing the CMOO problems in LTE scheduling. First, the MU function can be designed in such a way that different performance criteria are considered in the scheduling rule computation (MUSI). The optimal or even the near-optimal throughput regions under the satisfied QoS are not guaranteed under this approach. Therefore, some methods adopt the decoupling TDPS/FDPS architecture in order to distribute the scheduling objectives on different domains. For example TDPS can adopt different rules for the QoS satisfaction and the FDPS domain manages the system throughput and user fairness tradeoff. That is, the first stage prioritizes those flows approaching to the QoS requirement deadlines, whereas the second stage is in charge of the fairly allocation of the selected flows. Therefore, two different techniques are involved in the SSR based CMOO scheduling such as: the coupled TDPS/FDPS-SSR based CMOO and the decoupled TDPS-SSR/FDPS-SSR based CMOO.

Another method of CMOO is the dynamic based scheduling scheme DSR-CMOO. This is the novel approach and its role is to apply at each TTI, the best scheduling rule in order to maximize the weighted sum of each objective reward. The proposed scheme addresses at each TTI different objectives and aims to reach

and to keep as long as possible the optimality region with reduced system complexity. DSR-CMOO problems can be implemented in both coupled and decoupled TDPS/FDPS modes and the details about these concepts are discussed in Sub-section 8.5.2 in Chapter 8.

Addressing the QoS objectives simultaneously is not a new approach. For example, in [136], the throughput, delay and packet drop are adjusted concurrently. The QoS parameters are mapped in the unity cube that represents the QoS states of each flow. Each QoS parameter is normalized over the QoS requirements. Then, a flow which is mapped inside of the unity cube is considered to be satisfied from the viewpoints of all QoS requirements. The desired point is represented by the Cartesian point of (1, 1, 1). Different algorithms are proposed in [136] based on the distance of each QoS state for each flow to the desired Cartesian point. However, the proposals are addressed for the real-time applications which run on sensor nodes [136]. The proposed LTE scheduling method with RL addresses the exact situation with the amendment that, instead of having the unity cube, the algorithm maps each objective in particular reward functions. Of course, when all the objectives states are in the desired Cartesian point of (1,1,1..), then the reward values are maximized. This is the case of the proposed coupled TDPS/FDPS-DSR scheme which treats the CMOO for the system throughput and Jain Fairness tradeoff control [137]. Users are grouped on three classes based on the CQI reports. The state space is computed by using a 5-dimensional state space: normalized total cell throughput, JFI and the percentage of users located in different classes. The action set contains different discrete values for the GPF-SP parameter (α) which is adapted TTI-by-TTI. The reward function permits to set different tradeoff levels between the normalized throughput and the JFI value for each CQI class. The Q-learning approach is used as a RL algorithm and the MLPNN function is used to approximate the optimal parameterization steps for each scheduler state. Simulation results indicate that for different tradeoff levels, the proposed policies are able to achieve the optimal throughput region while maintaining a desired fairness between users from the viewpoint of the JFI quantitative measure.

The coupled TDPS/FDPS-SSR based CMOO is addressed in [56] being focused on the PLR and HoL packet delay objectives. The MU is designed as a function of $R_i^{PL}[t], \overline{R}_i^{PL}[t], d_i^{HoL}[t]$ and $\overline{d}_i^{HoL}[t]$. The scheduling rule being entitled Opportunistic Packet Loss Fair (OPLF) was introduced in Equation 3.44, in Sub-section 3.5.6. According to [56], the proposed scheme is able to outperform the PF, M-LWDF and GPF-PLF scheduling rules from the viewpoints of the system throughput, fairness, PLR and the HoL delay.

The tolerable average absolute deviation of transmission rate (AADTR) parameter can be used to control the fluctuations in transmission rates. A very interesting CMOO approach focusing on GBR and AADTR is proposed in [138], in which the original optimization problem of rates assignment and the non-convex constraint set is solved by using the dual optimization technique with the projection stochastic sub-gradient method [138]. The scheduler complexity grows by adding the processing of two Lagrange multipliers for the AADTR and GBR constraints. However, simulation results show that the analyzed method performs better than the GPF-MLWDF rule from the perspective of the system throughput and PDR performances while maintaining desired GBR and AADTR levels.

Modified versions of GPF-MLWDF and GPF-EXP1 rules based on virtual token mechanism are proposed in [139]. The idea is to use virtual tokens (VT) to change the rule representation from $d_i^{HoL}[t]$ to the HoL token delay in order to treat the SSR-CMOO problem focusing on GBR and HoL delay. The M-LWDF-VT and EXP1-VT decreases the PLR rate when compared with the conventional schemes while guaranteeing the required bit rate for larger VoIP and video flows when compared with the traditional GPF-MLWDF and GPF-EXP1 rules when the considered fairness parameters are $(\alpha = 1, \beta = 1)$.

The bankruptcy game [140] and Shapley value (SH) [141] as cooperative game theory are used in [142] to build a coalition between different flow classes in the TDPS stage in order to distribute the resources fairly. After the flow classification, the FDPS uses the GPF-EXP2 rule to improve the delay performance. The EXP2-SH improves the fairness and PLR indicators when the

CBR, VoIP and video traffic types are mixed. The same proposal can be mixed with the VT approach in the FDPS stage [143]. Alongside of the fairness and PLR improvement, the GBR target is also addressed due to the token queue representation. In the mixed VoIP and video scenario, the best option for video traffic is the EXP2-VT-SH rule, whereas for VoIP flows, the best performances are achieved by using the EXP2-SH scheduling scheme.

The performance of the LTE packet scheduling optimization for VoIP and BE mixed traffic is studied in [144] under decoupled TDPS/FDPS scheme. The TDPS scheduling is performed based on two concepts: required activity (RA) metric and delay sensitivity (DS) function. The RA function is calculated based on PFSch with GBR constraints $\left(\breve{T}_i[t]/\overline{T}_i^{Sch}[t]\right)$ in the time domain and the DS function takes different approaches depending on the $d_i^{HoL}[t]$ parameter (DS=1 for BE traffic). The proposed DS functions denote different degrees of prioritization for the VoIP traffic. Therefore, the TDPS stage is used to satisfy the rate and delay requirements. In the FDPS domain, the PFSch rule is performed in order to assign the RBs to the corresponding flows selected in the TDPS stage. Based on this strategy, it is shown that up to 346 VoIP users can be supported for the 5MHz system bandwidth. For the scenario with 200 VoIP and BE flows, the cell throughput is degraded by about 18% when compared with the full BE traffic scenario, when the soft prioritization for the VoIP flows is used.

An enhanced version of PF (E-PF) scheme for QoS guaranteeing is proposed in [145]. In the TDPS stage, users are grouped based on the CQI feedbacks and GBR requirements. Then, the FDPS schedules the classified users based on the ratio of the queue lengths and based on the ratio of RB air-time usage. The E-PF scheme is able to outperform PF from the viewpoints of the system throughput, mean packet delay and user fairness when the Poisson distribution is used for the traffic generation. Other approaches refer to the possibility of the user selection based on some priority values. If the priority is higher than a given threshold, the corresponding flow is selected. In [146], the fuzzy inference priority threshold generator is used to adjust adaptively the user

priority threshold at each time. In [147], the priority value is given by a novel time-utility function as a scheduling urgency factor for different traffic types.

The calculation of the necessary radio resources for the RT traffic while guaranteeing the QoS satisfaction represents one way to balance the tradeoff between QoS and the multi-user diversity. By adding RT traffic types with aggressive QoS requirements will lead to the loss in the opportunism. In order to minimize the opportunism loss, minimum resources which are necessary for the RT traffic should be determined. In [148] this principle is achieved by calculating the minimum rate for each user based on it's time to expire value. The noble γ parameter is used here to increase the importance of time to expire value for each packet. Basically, γ adjusts the tradeoff between the QoS importance and the multi-user diversity. The scheduling is divided in two parts: in the first part, the RT flows are scheduled based on PFSch rule, where $\breve{T}_i[t]$ is replaced by the newest minimum required bit rate at each TTI; and in the second part, the NRT traffic type is scheduled based on the remaining resources. By adjusting the parameters α, β and γ for the GPFSch-DP scheduling scheme, efficient tradeoff levels can be obtained between the QoS satisfaction, the throughput maximization and the user fairness assurance.

A.8 Summary

The main drawback of these methodologies discussed so far refers to the fact that the scheduling performances are not studied in terms of the optimality of different objectives at each TTI. The proposed scheduling scheme aims to increase the number of feasible TTIs when different DSR-SMOO/CMOO problems are solved by focusing on different scheduling objectives when the reinforcement learning approaches are used. In the following, the scheduling techniques are presented based on the addressed SMOO/CMOO combinatorial problems being categorized according to the following elements: scheduling technique, assumptions, analytical methods, network topology, the state space representation and the main focus of the considered proposals.

Table A.1 LTE Scheduling Strategies Based on SMOO

Scheduling Technique	Assumptions	Analytical Methods	Network Topology	State Space	Focus
Multiuser Scheduling on the LTE Downlink with Simulated Annealing [73]	-Limited CQI feedback: EESM SINR Mapping -Error Free Transmission -Best Effort Traffic Type -Infinite buffer	-Non-linear Optimization Programming -Simulated Annealing	-Fixed Power Allocation -Single Cell Downlink Transmission	-Achievable user rates	-Main Focus: Near-Optimal System Throughput -Other Focus: Complexity Overhead
Multiuser Scheduling on the in LTE Downlink with Meta-Heuristics [74]	-Limited CQI feedback: EESM SINR Mapping -Error Free Transmission -Best Effort Traffic Type -Infinite buffer	-Non-linear Optimization Programming -Genetic Algorithm	-Fixed Power Allocation -Single Cell Downlink Transmission	-Achievable user rates	-Main Focus: Near optimal System Throughput -Other Focus: Complexity Overhead
Joint Scheduling and Resource Allocation via ACK/NACK Feedback [75]	-Best Effort Traffic Type -Infinite buffer -Error Transmission -SISO Transmission	-Mixed Integer Greedy Linear Optimization with Lagrangian Relaxation -POMDP -Causal Global Genie -Particle Filtering	-Uniform Power Allocation -Single Cell Downlink Transmission	-ACK/NACK feedbacks - CQI based squared gain posterior distribution	Main Focus: System Throughput Other Focus: System Complexity
Exploiting Channel Memory for Joint Estimation and Scheduling [76]	-Non-full CQI feedback -Markovian channel model -Transmission with errors	-Markov Chain -Restless Multi-Armed Bandit Process [77] -Whittle's Index Policy [77] -Greedy Policy	-Single Cell Downlink Transmission	-Achievable User Rates -average user throughput -CQI state space -channel estimator and rate adapter pair -ACK/NACK feedbacks	Main Focus: Near-optimal Throughput
Throughput Maximizing Multiuser Scheduling with Adjustable Fairness [79]	-Full CQI Feedback -Error Free Transmission -Infinite Buffer -Best Effort Traffic	-DSR based SMOO -Probability Mass Function -Jain Fairness Index [80]	-Fixed Power Allocation -Single Cell Downlink Transmission	-Achievable User Rates -Average User Rates -Alpha Parameter -Jain Fairness Index	Main Focus: User Fairness Other Focus: Near-Optimal Throughput
Adaptive Fairness Control Proportional Fair [82]	-Best 5 CQIs values -Error Free Transmission -Infinite Buffer -Best Effort Traffic with GBR Requirements	-DSR-TDPS/FDPS -Linear Mean Square Approximation -Cumulative Distribution Function	-Fixed Power Allocation -Rayleigh Fading (Vehicular A) -Single Cell Downlink Transmission	-Achievable User Rates -Average User Rates -CDF and NGMN Req. Distance -Token Counter	Main Focus: User fairness when GBR satisfaction is fulfilled Other Focus: Near-Optimal Throughput

Fairness and Throughput Analysis for GPF Frequency Scheduling [85]	-Periodic and Full-CQI Reports -Error Transmission - Best Effort Traffic with Full Buffer Model	-Coupled TDPS/FDPS-SSR	-Single Cell Downlink Transmission -Fast Fading Channel: Jakes Model	-Achievable User Rates -Average User Rates - Alpha and Beta Parameters	Main Focus: User Fairness Other Focus: Optimal Throughput
Fair Weights for Heterogeneous Traffic Scheduling [88]	-Imperfect CQI Reports -Heterogeneous and Homogeneous Traffic Types -Error Transmission	-Coupled TDPS/FDPS-DSR -Lagrange Dual Decomposition [89] -Interior Point algorithm [90]	-Single Cell Downlink Transmission -Rayleigh Distributed Multipath Signals	-Achievable User Rates -Average User Rates -Deviation from Fair Weights, Fair Weights -Fairness Index	Main Focus: Utility Proportional Fairness Other Focus: System Throughput
Self-Organized Resource Allocation with Weighted Proportional Fair [91]	-Average, Perfect, Fast and Slow CQI Feedbacks -Best Effort Traffic	-Coupled TDPS/FDPS-DSR -Cross-Layer Distributed Protocol -Online Scheduling Policy -Selfish Strategy	-Multi-cell FDD Transmission - Rayleigh Fast Fading	- Achievable User Rates -Average User Rates -WPF Fair Weights -Average CQIs -Proportion of TTIs for Scheduled UEs	Main Focus: User Fairness Other Focus: System Throughput
Dynamic Packet Scheduling Performance in LTE [92]	-Periodic, Imperfect and non-full CQI Reports -Error Transmission -BE Traffic with Full Buffer	-Decoupled TDPS-SSR/ FDPS-SSR	-9 Macro-cell Sites with 3 Sector Antennas -FDD Transmission -Typical Urban (TU) Channel Type	-Achievable User Rates -Average User Rates -Average Achievable User Rates	Main Focus: User Fairness Other Focus: Near-Optimal Throughput
QoS Oriented Time and Frequency Domain Scheduler [93]	- Periodic, Imperfect and non-full CQI Reports -Error Transmission -Finite buffer size with Poisson Call Arrival -BE and CBR Mixed Traffic Types	-Priority Set Scheduler -CMOO based on Decoupled TDPS-DSR/FDPS-SSR -DSR: MaxFair., PF SSR:PF, TTA, PFSch	-Multi-cell Downlink Transmission -1x2 SIMO configuration -HARQ Ideal Chase Combining	-Achievable User Rates -Average User Throughput -Average Achievable User Rates -Average Scheduled User Throughput [94]	Main Focus: GBR satisfaction and user fairness Other Focus: Near-Optimal Throughput

Opportunistic Scheduling Scheme with Minimum Data-Rate Guarantees [96]	-Full and Perfect CQI Reports -Best Effort Traffic Type -Fully Back-logged Queues	-Heuristic-Coupled TDPS/FDPS-SSR -Determines the Maximum Number of Allowable RBs for each UE at each TTI	-Downlink Single Cell and TDD Transmission -Urban Macro Cell Scenario	-CQIs -Average User Rates -Average Achievable User Rates -Maximum Allowable RBs for each UE	Main Focus: GBR satisfaction Other Focus: User Fairness and Near-Optimal System Throughput
Multi-Carrier Gradient Scheduler based on Minimum/Maximum Rates Constraints [83], [98]	-Full and Errorless CQI Estimation -Transmission Without Errors -Full Buffer Model	-Coupled TDPS/FDPS-SSR	-Downlink Single Cell Transmission -With/Without Large Scale Fading (Path-loss and Shadowing)	-Achievable User Rates -Average User Rates -Token Counter Values -GBR/MBR Constraints	Main Focus: GBR/MBR satisfaction Other Focus: User Fairness and Optimal Throughput
Barrier Function based Proportional Fair Scheduling [54], [99]	-Wideband, perfect CQI with Power Control -Transmission with Errors HARQ, RLC-AM -Streaming Video Traffic	-TDPS-SSR -Possible Implementation for Decoupled TDPS/FDPS: TDPS-BF-PF/FDPS-PF	-Multi-cell Downlink Transmission -Multi-Path Fading with 3GPP Typical Urban Channel Type	- Achievable User Rates -Average User Rates -Average User Rate based on TWGR -GBR Constraints	Main Focus: GBR Satisfaction Other Focus: User Fairness and Near-Optimal Throughput
QoS-Aware PF Scheduling with Required Activity Detection [94]	-Error CQI Reports -Transmission with Errors -RAC Mechanism -Traffic Model: Single Packet	-TDPS-SSR	-HSDPA Macro-Cell Downlink Transmission -AWGN Channel	-Achievable User Rates -Average User Rates -Average Scheduled User Throughput - GBR Constraints	Main Focus: GBR Satisfaction Other Focus: User Fairness and Near-Optimal Throughput
Throughput Guarantee based on Violation Probability [103], [104]	-Perfect [103] and Imperfect [104] CQI reports -BE Traffic -Backlogged Queues	- Coupled TDPS/FDPS-DSR -Rare Event -Large Deviation Theory -Maximum A Posteriori Predictor [104]	- Single Cell Downlink Transmission -Channel with flat fading and AWGN Gaussian noise	-Achievable User Rates -Achievable User Rates Weights -Average User Throughput in TWRG -GBR Requirements	Main Focus: GBR Satisfaction Other Focus: User Fairness and Optimal Throughput
Modified-Largest Weighted Delay First [55], [106]	-Imperfect CQI Report -Transmission with Errors (H-ARQ) -Finite Buffer Size -Video Streaming based on CBR Traffic Type	- TDPS-SSR	-HSDPA Single Cell Downlink Transmission -ITU Pedestrian A Channel Type	-Achievable User Rates -Average User Rates -HoL Constraints -Instantaneous HOL Delay -Packet Dropped Rate	Main Focus: HoL Packet Delay Other Focus: User Fairness and Optimal Throughput

Exponential Proportional Fair [108], [109]	-Perfect CQI Reports -Infinite Buffer Type -Video Streaming Service Type	-Coupled TDPS/FDPS-SSR	-Single Cell Downlink Transmission	-Achievable User Rates -Average User Rates -HoL Constraints -Instantaneous HOL Delay -Packet Loss Rate Constraint -Average Instantaneous HoL for all UEs	Main Focus: HoL Delay Other Focus: User Fairness and System Throughput
Downlink Scheduling for Multiclass Traffic in LTE [110] (LOG-PF and EXP-PF Rules)	-Transmissions with Errors -Full Buffer Model -Heterogeneous QoS Constraints	-Coupled TDPS/FDPS-SSR	-Multi-cell Downlink Transmission -Modified HATA Urban Propagation Model [110]	-Achievable User Rates -Average User Rates -HoL Constraints -Instantaneous HOL Delay for each UEi -Average Instantaneous Delay for each UEp ($p \neq i$)	Main Focus: HoL packet Delay Other Focus: Queue Stability, User Fairness and Optimal Throughput
Modified-Earliest Due to Date Scheduling Rule [114]	-Perfect CQI Feedback -VoIP and Streaming Video Traffic Types	-Coupled TDPS/FDPS-SSR	-Single Cell FDD Downlink Transmission -Jakes Multipath Propagation Model	-Achievable User Rates -Average User Rates -HoL Constraints -Instantaneous HOL Delay	Main Focus: Packet Delay Other Focus: Packet Loss Rate, User Fairness, Optimal Throughput
Delay-Aware Packet Scheduling for Multiple Traffic Classes [115]	-Perfect CQI Feedback -FTP, Web, Video and VoIP traffic types	-Coupled TDPS/FDPS-SSR	-Single Cell Downlink Transmission -Modified COST-231 Hata Channel Model	-Achievable User Rates -Average User Rates -HoL Constraints -HOL Delays	Main Focus: HoL Delay Other Focus: User Fairness and Optimal Throughput
Frame Level Scheduler for Real Time Services [117], [118]	-Full, Errorless and Periodic CQI Feedback -RT Traffic: H264, VoIP -BE Traffic: Infinite Buffer	-Decoupled TDPS/FDPS (TDPS-FL/FDPS-SSR) -Frame Level Scheduling based TDPS -FDPS-MT or PF	-Multi-cell FDD Transmission with Interference -Rayleigh Fading Channel Model	-Achievable User Rates -Average User Rates -HoL Delay Constraints -Instantaneous Queue Size	Main Focus: HoL Delay Other Focus: User Fairness and System Throughput

Utility Based Resource Allocation Scheme with Delay Scheduler [119]	-Full, Errorless and Periodic CQI Feedback GBR Traffic: VoIP Non-GBR: Video, Gaming, CBR.	-Decoupled TDPS/FDPS -TDPS: Game Theory, Delay based Utility, Lagrange Decomposition FDPS: DPS Scheduling [120]	-Single Cell Downlink Transmission with Interference -Multi-path: Jakes Model	-Achievable User Rates -Average User Rates -HoL Delay Constraints -HOL Delay for Each Flow	Main Focus: HoL Delay for Heterogeneous Constraints Other Focus: PLR, Fairness and Optimal Throughput
Priority-Coupling-A-Semi-Persistent MAC Scheduler [123]	-Perfect and Errorless CQI Feedback -Traffic Type: VoIP	-Decoupled TDPS/FDPS -TDPS: Priority Mode based on Coupling Method -FDPS: CAFQ [124]	-Single Cell Downlink Transmission	-CQI feedbacks -Achievable User Rates -Queue Size -HoL Delay and PDR Constraints	Main Focus: HoL Delay for VoIP Users Other Focus: User Fairness and System Throughput
Intelligent Scheduling Architecture for Mixed Traffic focusing on Packet Loss Rate [127]	-Perfect CQI Feedback -RT Buffer: Poisson Arrival -NRT: Infinite Buffer	-Decoupled TDPS/FDPS -TDPS: Hebbian Learning [128] and K-Means Clustering [129] -FDPS:SSR (MT, PF)	-Single Cell Downlink Transmission -Channel: COST 231 Walfisch-Ikegami Model	-Achievable User Rates -Average User Rates -Instantaneous PLR and PLR Constraints -RBs Proportion	Main Focus: Packet Loss Rate Other Focus: HoL Delay, User Fairness and Near-Optimal Throughput
Max-Delay Utility Scheduling Rule based on Queue Stability [49], [50], [133], [134]	-Perfect and Periodic CQI Reports -Mixed Traffic Types: Streaming, Voice, BE.	-Coupled TDPS/FDPS	-Single Cell Downlink Transmission -Multi-Path Rayleigh Fading	-Achievable User Rates -Average User Rates -Queues Sizes -Average HoL Delays -Average Arrival Rates	Main Focus: Queue Stability, Maximum Stability Region Other Focus: HoL Delay, User Fairness, Optimal Throughput

Table A.2 LTE Scheduling Strategies Based on CMOO

Scheduling Technique	Assumptions	Analytical Methods	Network Topology	State Space	Focus
A Novel Dynamic Q-Learning-Based Scheduler Technique with Neural Networks [137]	-Full and Period CQI Feedback -Infinite buffer -BE Traffic Type -Transmission with Errors -Equal Time for Exploration and Exploitation	-GPF-SP-TDPS/TDPS -Q-Learning Algorithm for GPF-SP Parameter Selection -State Space Approximation: Single Layer Perceptron Feed-forward Backward Neural Network	-Single Cell Downlink Transmission -Multi-path: Jakes Model	-Achievable User Rates -Average User Rates -Percentage of Users with Poor, Medium and Good Channel Conditions -Normalized System Throughput -Jain Fairness Index	Focus: Optimal System Throughput and Jain Fairness Index
Opportunistic Packet Loss Fair Scheduling for Delay Sensitive Applications [56]	-Full and Error Free CQI Feedback -Arrival Rate: Truncated Pareto Distribution -Traffic Type: Video Streaming	-Coupled TDPS/FDPS-SSR	-Single Cell Downlink Transmission -Channel Type: 3GPP Typical Urban -Path-Loss Model: Hata-Cost-231	-Achievable User Rates -Average User Rates -instantaneous HoL Delay and PLR -HoL and PLR Requirements	Main Focus: HoL delay and Packet Loss Rate Other Focus: System Throughput and User Fairness
Resource Allocation in OFDMA Wireless Communications Systems with Multimedia Services [138]	-Queue Model :CBR for RT and Infinite for BE	-Coupled TDPS/FDPS-SSR -SSR: Dual Optimization Technique and Projection Stochastic Sub-gradient Method	-Single Cell Downlink Transmission -Channel Model with Uncorrelated Scattering -Multi-Path with Path Loss and Shadowing	-Achievable User Rates -Average User Rates -GBR and AADTR Constraints -Lagrange Multiplier for GBR and AADTR	Main Focus: GBR and Average Absolute Deviation of Transmission Rate (AADTR) Other Focus: Throughput and User Fairness
M-LWDF and EXP/PF Based on Virtual Token Mechanism [139]	-Perfect CQI Feedback -Traffic Type: 40% VoIP, 40% Video and 20% FTP -Video: Arrival Rate with 242 kbps -VoIP: G.729 voice Flow -FTP: Infinite Buffer	-Coupled TDPS/FDPS-SSR -SSR: Instead of Using the HoL Delay, a HoL Token Delay is Considered	-Single Cell Downlink Transmission with Interference -MultiPath Loss: Jakes Model	-Achievable User Rates -Average User Rates -HoL Delays -Token Queue Size -Average Instantaneous HoL Delays for all UEs -GBR Constraints	Main Focus: HoL Delay and GBR satisfaction Other Focus: User Fairness and Optimal Throughput

Resource Allocation Using Shapley Value [142]	-Perfect CQI Feedback -Traffic Type: 40% VoIP, 40% Video and 20% CBR	-Decoupled TDPS/FDPS -TDPS: The Bankruptcy Game [140] and Shapley Value [141] (Cooperative Game Theory) -FDPS:EXP-RULE	-Single Cell Downlink Transmission with Interference -Multi-Path Loss: Jakes Model	-Achievable User Rates -Average User Rates -HoL Delays -Average Instantaneous HoL Delay for each UEp (p≠i) -Shapley Values	Main Focus: User Fairness and HoL Delay Other Focus: GBR Satisfaction and Optimal Throughput
Resource Allocation Using Cooperative Game Theory and Virtual Token Mechanism [143]	-Perfect CQI Feedback -Traffic Type: 50% VoIP, 50% Video.	-Decoupled TDPS/FDPS -TDPS: The Bankruptcy Game [140] and Shapley Value [141] (Cooperative Game Theory) -FDPS:EXP-RULE with or without Virtual Token Queues.	-Single Cell Downlink Transmission with Interference -Multi-Path Loss: Jakes Model	-Achievable User Rates -Average User Rates -HoL Delays -Average Instantaneous HoL Delay for each UEp (p≠i) -Shapley Values -Token Queue Size -Average Instantaneous HoL Delays for all UEs -GBR Constraints	Main focus: User Fairness, GBR Satisfaction and HoL Delay Other Focus: Optimal Throughput
Dynamic Packet Scheduling for Traffic Mixes of BE and VoIP Users [144]	-Special RBs Assignments for Queue Stability -CQI Feedback with Delay and Errors -Traffic Types: Mixes of VoIP and BE	-Decoupled TDPS/FDPS -TDPS: SSR focused on GBR Requirement and HoL Packet Delay -FDPS:SSR based on Simple PFSch	-Single Cell Downlink Transmission -Antenna Configuration: 1TX, 2RX -HARQ Model: Chase Combining	-Achievable User Rates -Average Scheduled User Rates -GBR and HoL Delay Constraints -Instantaneous HoL Delay	Main Focus: User Fairness, GBR, HoL Delay Satisfaction and Queue Stability Other Focus: Near-Optimal Throughput
An Enhanced Proportional Fair Scheduling for QoS Guarantee [145]	-Full and errorless CQI reports -Traffic type: Poison arrival distribution	-Decoupled TDPS/FDPS -TDPS: Users are classified based on the CQI reports and based on the GBR priority -FDPS: SSR depending on queue length ratio and RB-airtime usage ratio	-Single Cell Downlink Transmission	-Achievable User Rates -Average User Rates -GBR Constraints -CQI Feedback -Queue Length Ratio -Resource Blocks Airtime Usage Ratio	Main Focus: GBR, HoL Delay, Queue Stability and Fairness Satisfaction Other Focus: Optimal System Throughput

Appendix B

CQI Cycle in LTE Networks

B.1 Appendix Outline

The CQI report is crucial in LTE scheduling when the DSR-SMOO/ CMOO problems are approximated at each TTI based on the scheduler state space. The channel information can improve the sustainability of the proposed scheduling policies especially when the trade-off between system throughput and user fairness is analysed. The impact of the propagation loss model in the SINR generation is studied in the following. The CQI report is obtained by using some quantization methods from the obtained SINR levels.

B.2 Propagation Loss Modeling

Each step involved in the generation process of CQI feedbacks starting with the reference signal transmission is highlighted in Fig. B.1. The reference signals (pre-known to both user and eNodeB) are sent at each TTI by the eNodeB station over the whole system bandwidth. Then, this signal is attenuated by the propagation loss model and by the accumulated interferences from other cells that are using the same frequency range. Based on the reference signals, each user measures the channel gain or the signal-to-interference/noise ratio of each RB and

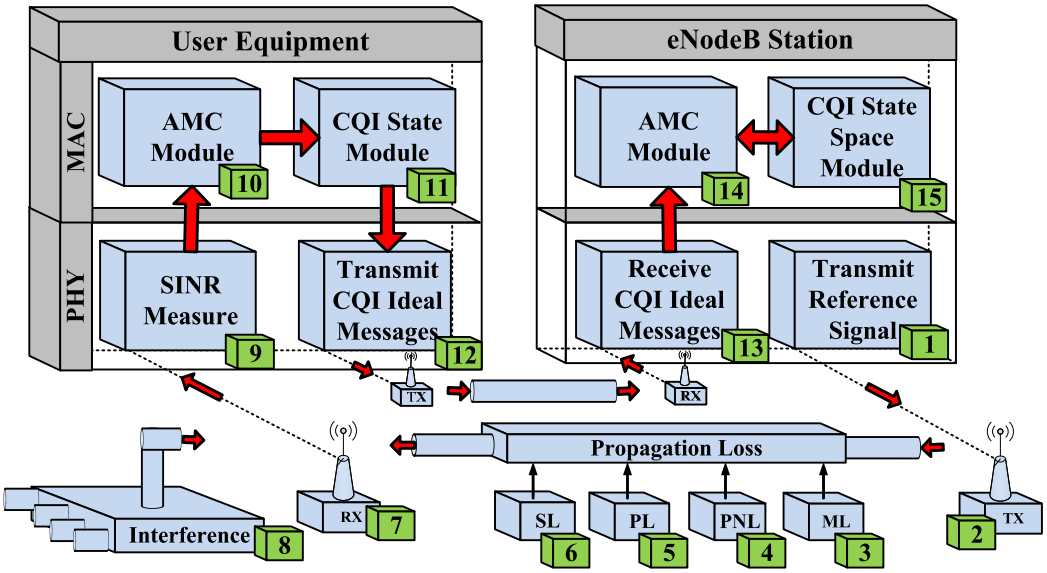
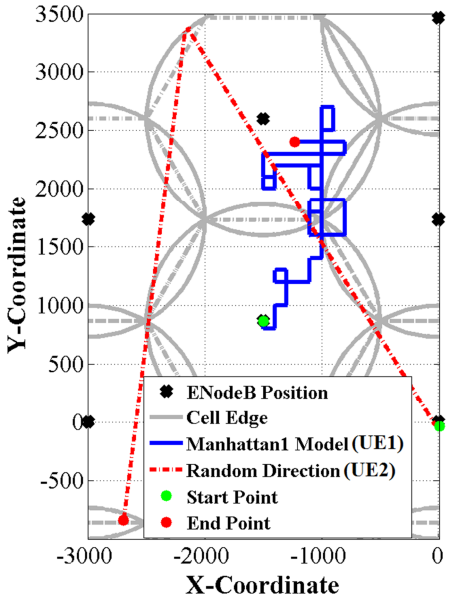


Fig. B.1 CQI Cycle in LTE Networks

converts in its quantized version of CQI value. The conversion is achieved at the UE MAC layer by using the functionalities of the AMC module. The channel state information (CQIs for each RB) is transmitted to the eNodeB via a separate channel feedback such as PUCCH. At the eNodeB side, the CQI message is considered to be received without errors, and this message is forwarded to the MAC layer where is analyzed by the AMC module and further provided to the CQI state space module. At this level, the CQI report for each active user is prepared for the classification and regression stages. For the rest of the simulations, the following assumptions are considered: the feedback channels are errorless, each RB has a flat fading for the whole sub-band and the SINR estimation is constant within one TTI.

The analyzed CQI cycle considers a macro-cell urban area scenario [151] with two users which experience two channel types [153], [154], [155] (Fig. B.2.a.): Jakes Model with Manhattan ground area and ITU Vehicular Type A with high-way random direction mobility, respectively [156]. The rest of parameters are presented in Fig. B.2.b imported from the 3GPP [36], [149].

The Downlink Reference Signal (DSR) is used to cope up different aspects of LTE systems such as: channel estimation and demodulation of the control information, measurements for CQI, RI and PMI, handover decision and cell



a)

Parameter Name	Value
Transmission Power (P^{TX})	43 dBm
Manhattan (UE1) Multipath	Jakes Model
Random Direction (UE2) Multipath	Vehicular A Model
Path Loss Model	Macro Cell Urban
Penetration Loss	10 dB
Noise Figure (F)	2.5
Noise Spectral Density (N_0)	-174 dBm
CQI Reporting Mode	Periodic
CQI Reporting Period	1ms
Number Of Cells	19
Frequency Reuse Factor	3
Number Of RBs	25
RB Bandwidth	180KHz
MS speed	30 Km/h
Simulation Time	1000 s

b)

Fig. B.2 Test Case Scenario a) Mobility Models b) List of System Parameters

selection [157], [158]. Hence, as suggested in [157], the DRS signal can be divided in two categories: cell specific and user specific reference signals. However, the SINR measurement for CQI calculation and handover decision is based on the cell specific reference signal which is common for all users. In Fig. B.3 is depicted the typical case with SISO model in which the reference signal is broadcasted. In the absence of other antenna configurations, the remaining resource elements are used by the PDCCH and PDSCH logical channels. The transmission power for the whole bandwidth is considered to be *43 dBm*.

In general the propagation loss models can be divided in three categories [159]: abstract propagation models, path loss aware models and fading loss aware models. The fading models increase the accuracy of the calculated propagation loss when compared with other categories due to the fact that each change in the propagation environment is considered. In order to minimize the computation complexity, the fading processes are modeled as stochastic realizations considering the following elements such as: *fast fading, shadowing loss, penetration loss and path loss* [156], [159].

Fading Models. The Jakes fading type is considered to be a deterministic model based on the Rayleigh fading in which the principle of summing the sinusoids is used [153]. It is important to notice that the Jakes model chooses the

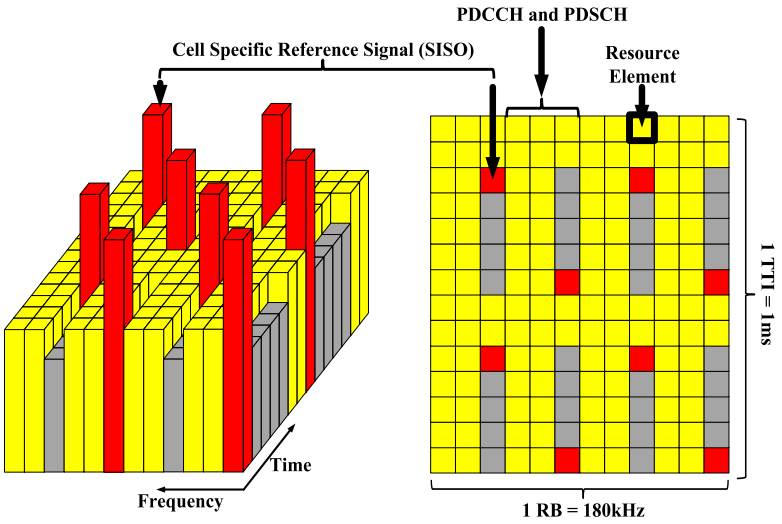


Fig. B.3 Cell-Specific Reference Signal (based on [157])

path gain, the initial phase and the Doppler frequency. Also, there is no cross correlation between the imaginary and the real parts of the modeled Rayleigh waveform. For the LTE fading process, the Jakes models consider a set of parameters such as: the central frequency of 2GHz and the system bandwidth in order to determine the periods of sinusoids, the user speed to determine the pulsation and the number of paths for the initial phase calculation. For this example, six paths are randomly generated as implemented in [156]. The time-frequency representation of multipath propagation for UE1 is depicted in Fig. B.4 and it corresponds to Point 3 from Fig. B.1.

The Zheng model is a non-deterministic model that proposes to reintroduce the randomness in the parameter selection in order to simplify the reference model while assuring a higher order for the statistical properties [160] (Fig. B.5). The ITU Vehicular Type A channel model [150] is used to generate the channel gains for a realistic transmission scheme.

PeNetration Loss (PNL) (Point 4) considers the concrete wall attenuation and it is fixed to 10dB for the frequency range of 2GHz as described in [155].

Path Loss (PL) (Point 5) represents the most important factor in the signal attenuation which depends on many factors such as: free-space loss, reflection, diffraction. In LTE, the path loss parameters depend on outdoor cell model and the distance between eNodeB and UE as indicated in [151], [161].

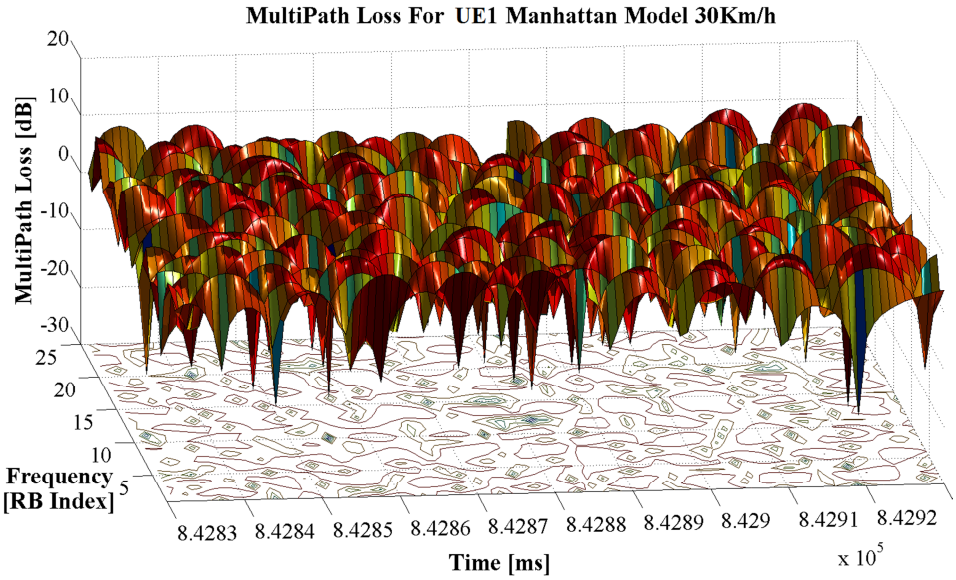


Fig. B.4 Multi-Path Loss for Jakes Model (Point: 3)

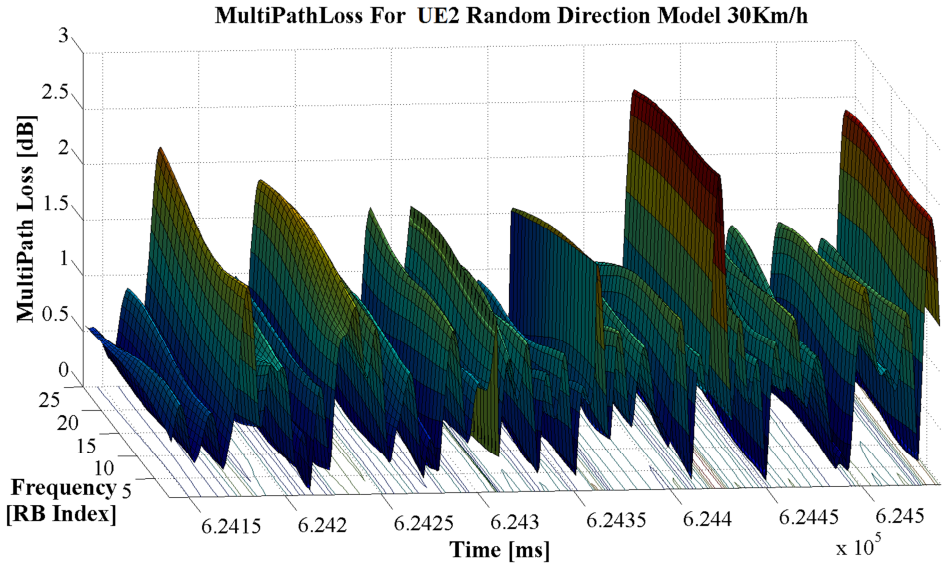
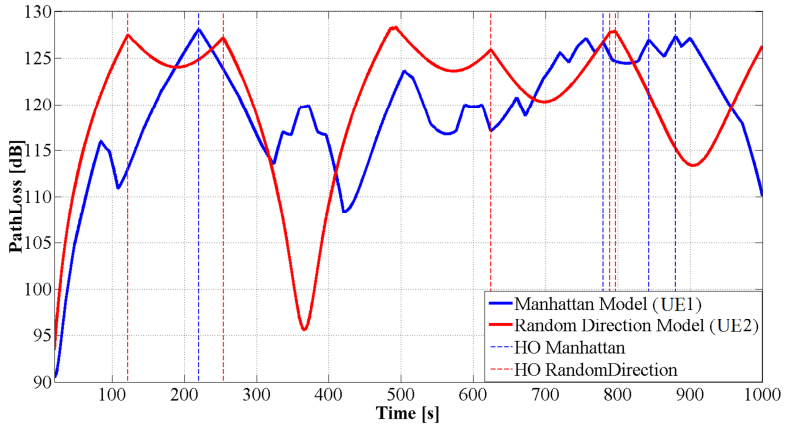


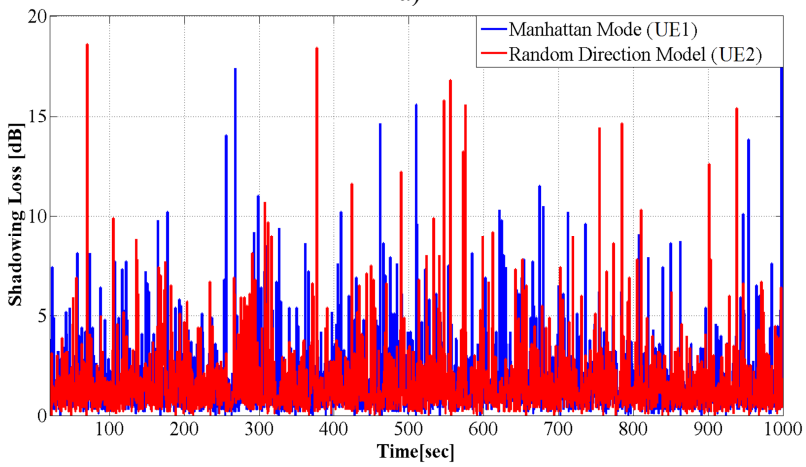
Fig. B.5 Multi-Path Loss for ITU Vehicular A (Point: 3)

Shadowing loss (SL) (Point 6) depends mainly on the obstructions that can appear between eNodeB and UE. For this reason, this parameter is modeled as a random process with a log-normal distribution. Figure B.6.a shows the path losses with the points where the handover is requested, and different shadowing loss values are created in the range of [0, 20] dB (Fig. B.6.b) for $(\mu = 0, \sigma = 8dB)$.

Received power (Point 7) is calculated based on the transmission power and the propagation loss effect as described in the following:



a)



b)

Fig. B.6 a) Path-Loss and b) Shadowing Loss

$$P_{i,j}^{RX}[dB] = P_j^{TX} - ML_{i,j} - PNL_i - PL_i - SL_{i,j} \quad \text{(B.1)}$$

where P_j^{TX} is the transmission power, $ML_{i,j}$ is the multi-path loss, PNL_i is the penetration loss, PL_i and $SL_{i,j}$ are the path loss and the shadowing loss, respectively for UE i and RB j.

SINR Levels Estimation (Point 9) represents the ratio between the received power $P_{i,j}^{RX}$ and the noise and interferences for each UE i and for each RB j for a given LTE bandwidth. Mathematically, the SINR calculation at the PHY-UE layer is [151]: $SINR_{i,j}[dB] = P_{i,j}^{RX} - NI_i$,where $NI_i = F \cdot N_0 \cdot BW_{RB} + I_i^{Inter} + I_i^{Intra}$ with the internal noise power F (default value 2.5), with the noise spectral density N_0, with inter-cell I_i^{Inter} interference and with intra-cell I_i^{Intra} interference.

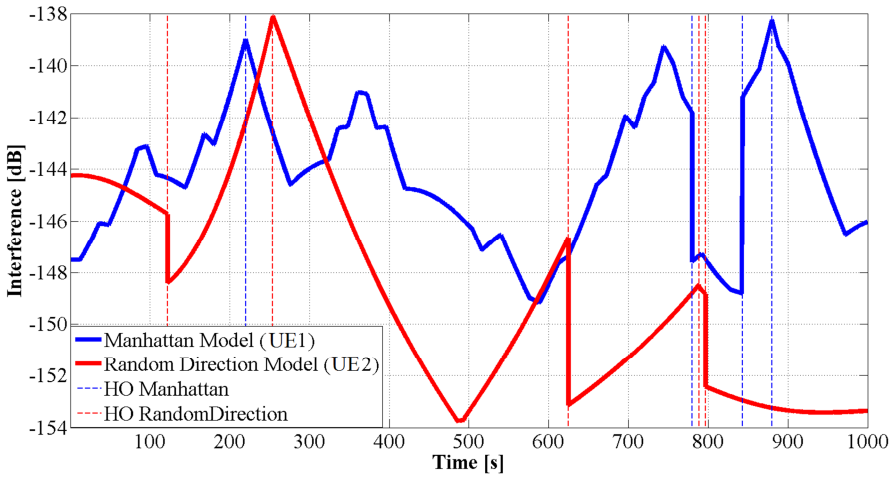


Fig. B.7 Interferences for Manhattan and Random Direction Mobility Models

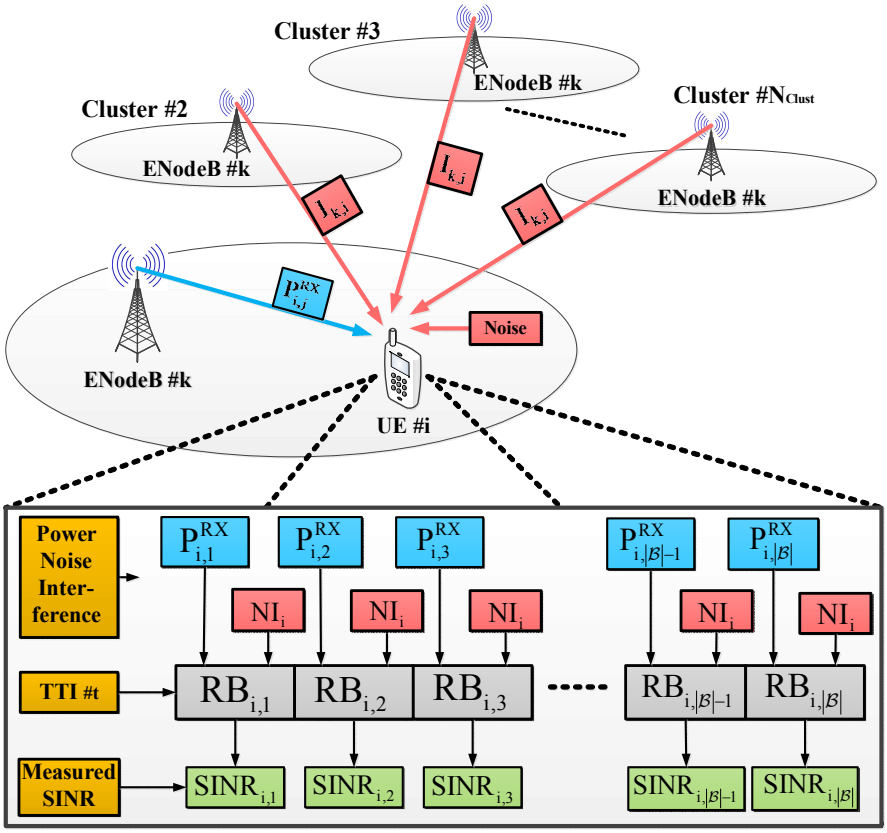


Fig. B.8 SINR Estimation Procedure

Interference Model: (Point 8) is based on the frequency reuse principle while keeping the transmission power to a constant level. The frequency range is divided by some neighboring cells where the interference could be very high. The set of cells that are dividing the frequency spectrum between them is called cluster

or site (Fig. B.8). The interference levels which are considered in this study are based on the neighboring cells of different clusters in which the same frequency spectrum is used. In this example, the considered number of cells is 19 and the frequency reuse factor is equal to 3 meaning that the scenario contains 6 clusters. Each cluster uses the downlink bandwidth of [2110, 2125] MHz. The SINR estimation procedure based on the considered interference model is depicted in Fig. B.8 and the interference levels for both users are highlighted in Fig. B.7. The highest difference in the interference level for UE1 of 9dB is represented by the moment of the third handover. In this point, the SINR level decreases due to the fact that the interference value increases. In the random direction mobility case (UE2), at the half part of the simulation period, the interference is minimized due to the fact that UE2 is located at the edge of the cell with less interference effects (without neighboring cells). The greatest variance of interferences is located at the 3rd handover time where the level of the interference decreases by about 6dB.

The impact of the interference in the obtained signal for both cases is illustrated in Fig. B.9.a and Fig. B.9.b. In the first case, the SINR level is seriously degraded due to the interference effect at the time of the third handover. In the second case, the reduction of the interference performance increases the SINR level by about 5dB when the third handover is considered for user UE2.

B.3 SINR to CQI Mapping Procedure

In the LTE system developments, there are two important aspects involved in the implementation process: the *algorithm implementation* and *testing and optimization procedures* for the overall network. In this sense, the simulations can be categorized in two parts:

1. ***Link-level simulations*** – in which particular segments of the simulator are implemented such as: fast fading generation, channel coding and decoding and general PHY processes;
2. ***System-level simulations*** – refers to the performance of the proposed algorithms to the entire cellular network. It is the case of the scheduling process or the case of handover decisions.

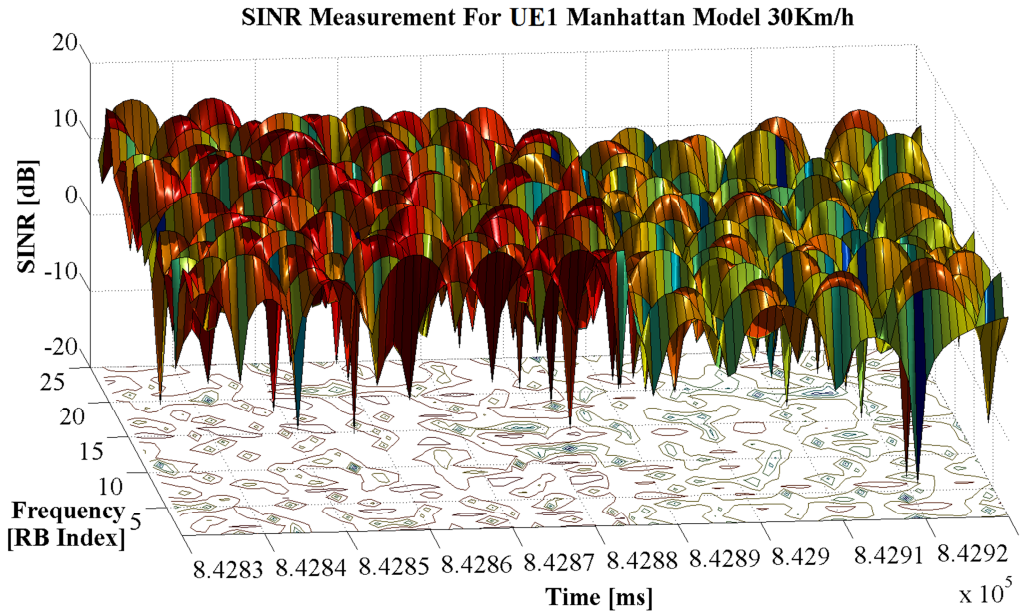


Fig. B.9 a) SINR Estimation for Jakes Model

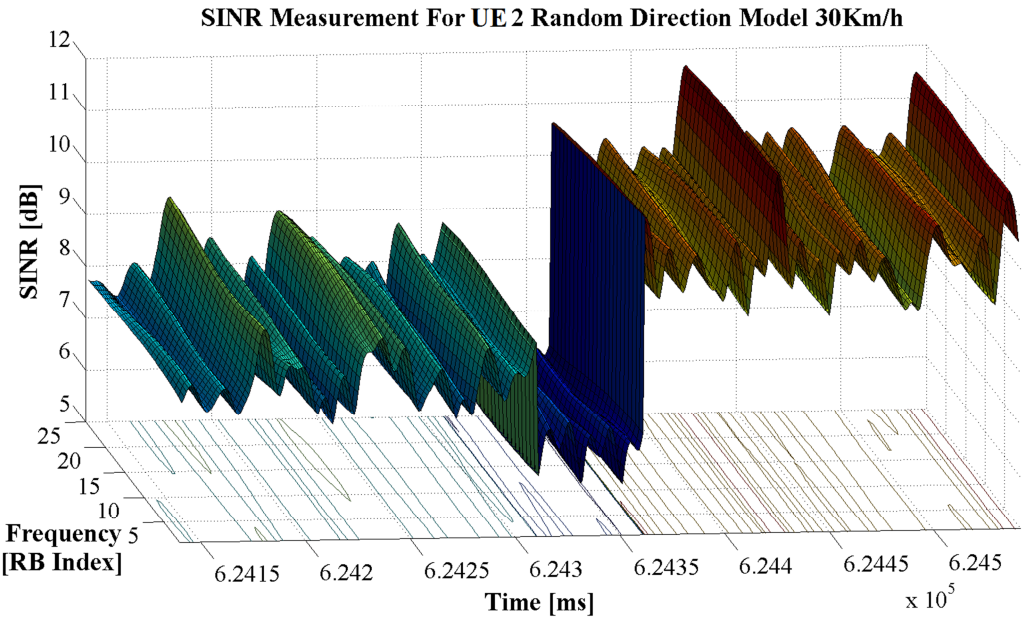


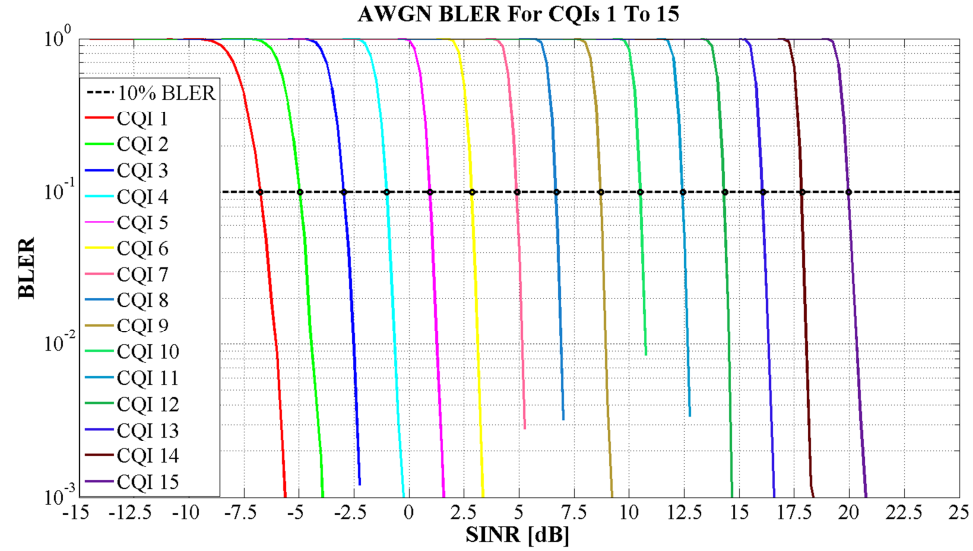
Fig. B.9 b) SINR Estimation for Vehicular A Model

The link level simulations lead to the high computationally cost. In this sense, an abstracting procedure of these processes is absolutely necessary. For instance, the fading modeling is based on the predefined traces. At the PHY level, the processes are abstracted sufficiently to produce high accuracy and reduced complexity. In this case, the mapping procedure between SINRs and CQIs is fulfilled based on the mapping curve which is obtained on the link-level

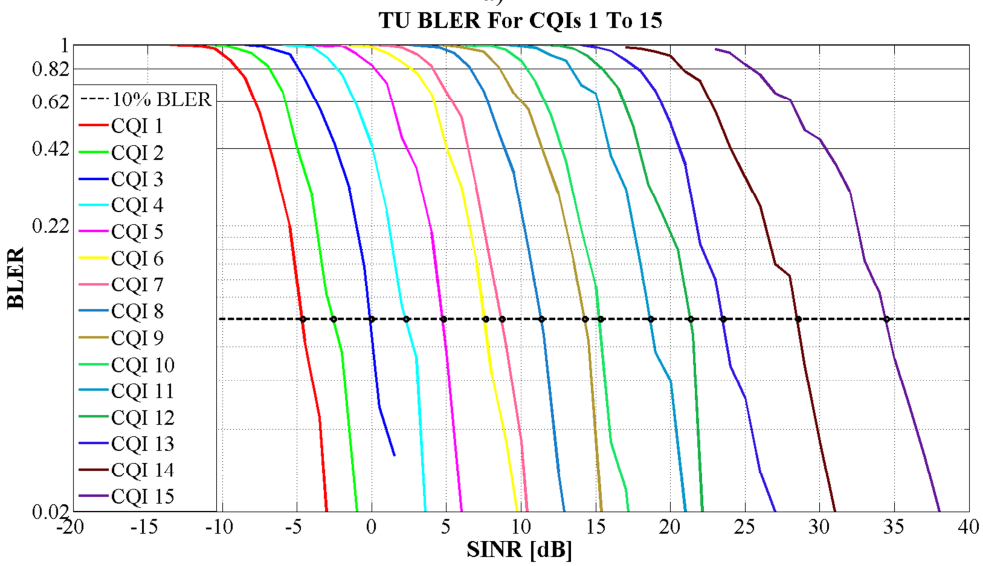
simulations. The discussion can be further extended to the supervised learning and reinforcement learning algorithms. The exploration and validation stages can be viewed as link-level simulations due to the fact that the precision and the accuracy results are considered to be crucial at these stages. In the exploitation stage, the performance of the entire network is analyzed based on the trained architecture.

At the link-level stage, the BLER values are determined by a given set of inputs at the receiver side such as SINR levels and different MCS schemes. In LTE networks, 30 MCS schemes are defined by containing the code rates between 1/13 and 1 and three modulations schemes such as QPSK, 16-QAM and 64-QAM [162]. Then, a set of 15 CQIs values are defined and subtracted from these data sets leading to different channel qualities from 1 to 15. In this way, the transmission overhead is reduced since the CQI report can be transmitted by using only 4 bits. Basically, BLER is calculated for each RB and represents the ratio between the number of TBs received with errors and the number of sent TBs [162]. In order to obtain the BLER value for each received TB, the link-level simulator proposed in [163] is used in the context of the current research. The SINR–BLER mapping process is achieved based on two sets of curves: AWGN and TU channels types. For each CQI curve, the SINR-BLER mapping procedure requires the SINR value for each RB. The SINR-BLER curves for SISO transmission with a bandwidth of 1.4 MHz and 5000 TTIs simulation time for both types of channels (AWGN and TU) are depicted in Figs. B.10.a and B.10.b.

After the scheduling decision, the TB is computed for each selected user and transmitted via the PDSCH channel. At the same time, the MCS scheme is associated to the TB size in order to inform the scheduled user what MCS scheme should be used for the decoding and demodulation processes. At the reception side, each user computes the BLER values based on the stored curves, by taking into account the measured SINR and the MCS received from the eNodeB. If the calculated BLER is less than the target BLER, then the TB is considered errorless and the ACK message is provided to the base station. Otherwise, a NACK message is sent together with the CQI report on the PUCCH channel. In Fig. B.10.a, if UE receives the MCS of CQI 15 to decode the data, and based on the



a)



b)

Fig. B.10 Reference BLER for a) AWGN and b) TU Channels

measured SINR, the BLER value is less than 10%, then the TB is considered erroneous and it needs to be retransmitted. For the considered simulations, the target BLER is considered to be set at 10% of the obtained BLER (stored curves).

Based on the target BLER, the reference values for SINR can be obtained for different channel types. That is, the intersection between the target BLER and BLER curves determines the labeled SINR values. The reference SINRs for both AWGN and TU channels are illustrated in Table B.1. For this scenario, the TU channel modeling is been used for the MCS decision. For instance, if the measured SINR is 15dB, then the selected modulation scheme is 64-QAM, the TB size is 328 and the corresponding CQI value is 10. In the general case of SISO

transmission on AWGN and TU channels, the mapping curve from the SINR levels to the CQI quantized values is illustrated in Fig. B.11. The spectral efficiency is based on the Shannon's theorem which is considering the channel capacity [164]: $\eta_{i,j} = \log_2\left(1 + SINR_{i,j} / \Gamma\right)$, where $\Gamma = -\ln(5 \cdot BLER)/1.5$ depends on the target BLER value and represents the SINR gap between practical and theoretical results. Based on the spectral efficiency calculated for each RB *j,* the CQI discrete values are obtained by using Table B.1.

By considering the proposed scenario, the CQI quantization process for both, Jakes and Zheng fading models in the time-frequency domain is illustrated in Fig. B.12.a and Fig. B.12.b, respectively. The impact of the handover procedure and the difference levels in the interference signals have a great impact in the CQI quantized values. For the Zheng model, the CQI vector can be classified in 15 classes due to the flat nature of the component elements in all cases. For the Jakes model, the variation of CQI elements requires more than 15 classes in order to have acceptable accuracy of the classified elements. The studied case uses the 5MHz bandwidth. In order to eliminate the number of RBs dependency and implicitly the system bandwidth dependency, the CQI vector has to be preprocessed first and then classified for a much better representation of the controller state space.

Table B.1 Mapping from a) SINR to CQI and b) Spectral Efficiency to CQI

CQI Index	Spectral Efficiency [bps/Hz]	Modula -tion Scheme	Transport block size [bits]	Reference SINR [dB] (AWGN)	Reference SINR [dB] (TU)
0					
1	≤ 0.15	QPSK	16	≤ -6.15	≤ -4.63
2	(0.15; 0.23]	QPSK	32	(-6.15; -4.37]	(-4.63; -2.6]
3	(0.23 0.38]	QPSK	56	(-4.37; -2.37]	(-2.6; -0.12]
4	(0.38 0.6]	QPSK	328	(-2.37; -0.42]	(-0.12; 2.26]
5	(0.6; 0.88]	QPSK	120	(-0.42; 1.53]	(2.26; 4.73]
6	(0.88; 1.18]	QPSK	136	(1.53; 3.43]	(4.73; 7.53]
7	(1.18; 1.48]	16-QAM	144	(3.43; 5.46]	(7.53; 8.67]
8	(1.48; 1.91]	16-QAM	208	(5.46; 7.25]	(8.67; 11.32]
9	(1.91; 2.41]	16-QAM	256	(7.25; 9.28]	(11.32; 14.24]
10	(2.41; 2.73]	64-QAM	328	(9.28; 11.11]	(14.24; 15.21]
11	(2.73; 3.32]	64-QAM	376	(11.11; 13]	(15.21; 18.63]
12	(3.32; 3.90]	64-QAM	440	(13; 14.9]	(18.63; 21.32]
13	(3.90; 4.52]	64-QAM	520	(14.9; 16.64]	(21.32; 23.47]
14	(4.52; 5.12]	64-QAM	584	(16.64; 18.41]	(23.47; 28.49]
15	≥ 5.12	64-QAM	712	(18.41; 20.54]	(28.49; 34.6]

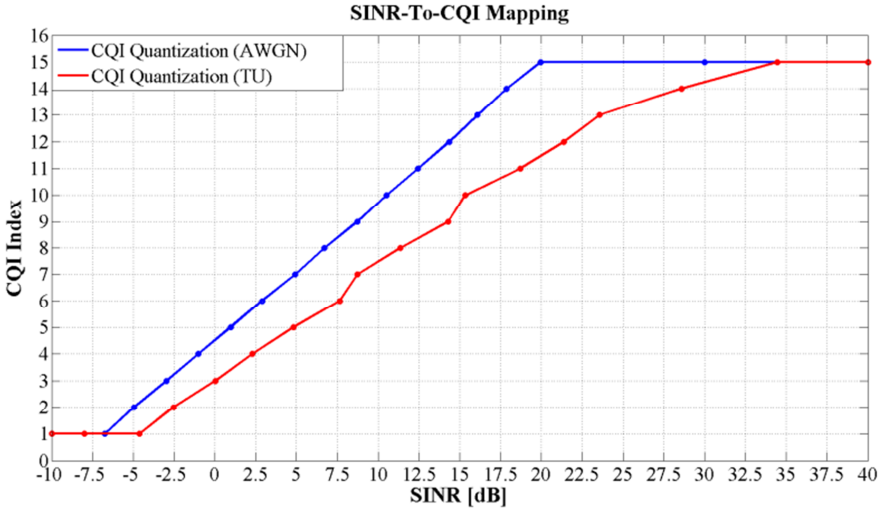
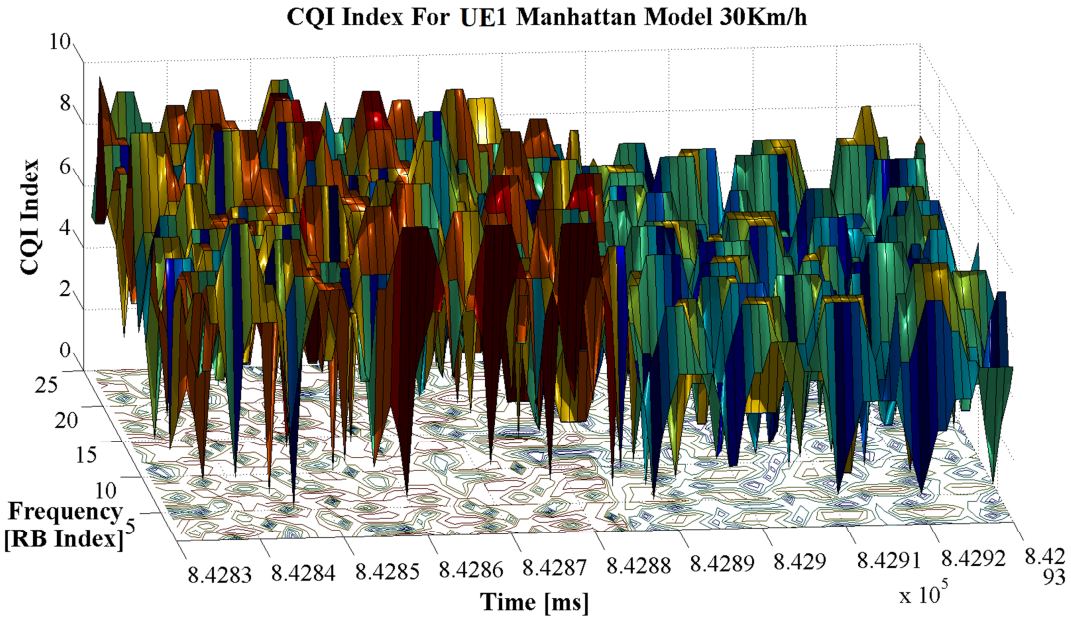
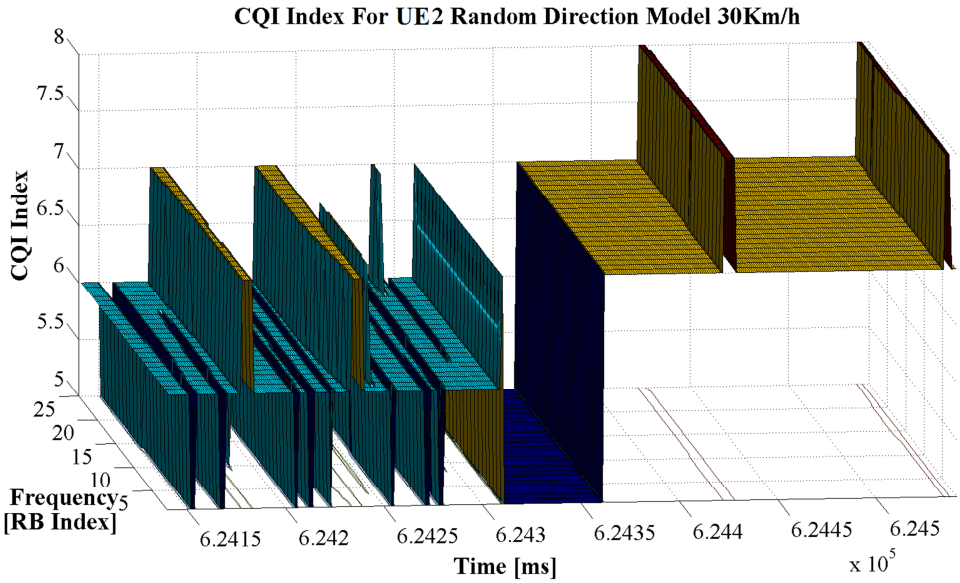


Fig. B.11 SINR to CQI Mapping Procedure

The proposed aggregation method of the uncontrolled CQI parameters aims to reduce the communication overhead for the CQI feedback message on the uplink side. Instead of reporting the quality for the whole bandwidth by using a proper clustering approach, each UE can report the closest center index for the corresponding report. On other hand, the training procedure of the RBFNN weights must be achieved under multiple preprocessed CQI observations. For this reason, the Zheng model is not suitable for the CQI classification procedure due to its flat nature. The Jakes fast fading model can provide other benefits for the supervised learning step such as the avoidance of the local minima solutions.



a)



b)

Fig. B.12 CQI Index for a) Jakes Model and b) Vehicular A (Zheng Model)

B.4 Summary

The cycle of the CQI report has been analyzed starting from the reference signals, propagation loss modeling and continuing with the quantization procedure from the SINR levels to the discrete CQI values. The Zheng fading type is not recommended to be used for the RBFNN classification due to its flat nature which can slow down the learning procedure and the trained structure can suffer from the local minima and under-fitting problems. The Jakes fast fading model provides higher variations of the CQI values among a given system bandwidth, and it is decided to be used for the entire set of simulations in this research due to its ability of improving the learning speed in the CQI classification procedure. Also, the local minima problems are avoided especially when the validation set is considered in the RBFNN training procedure.

Appendix C

Preprocessing Stage in CQI Aggregation

C.1 Appendix Outline

The preprocessing stage aims to reduce the dimension of CQI report for each active user to a more compressed dimension which depends on the number of CQI discrete values (15 in LTE). Even under this form, the size of the preprocessed CQI state space is very high for some bandwidths. In this sense, two mass modes are proposed in order to reduce much more the preprocessed state space size. The first method is the top mass principle which aims to select only the top CQI mass values which represent in fact the top number of RBs which is reporting the considered CQI discrete values. The second method is the majority mass mode which aims to select the CQI values being reported by a given majority percentage of the resource blocks. Based on these approaches, different CQI preprocessed mass mode schemes are analysed for each system bandwidth in LTE networks. The idea is to connect these methodologies to the CQI cycle module from Appendix B and to collect as many different preprocessed CQI observations as possible. In this direction, a novel collection algorithm is proposed. The collected sets of preprocessed mass observations are used for the clustering algorithms and serves as validation sets when the RBFNN structures are used in the classification of the preprocessed CQI observations.

C.2 The Initial Preprocessing Stage

The CQI-PS stage in LTE scheduling implies the translation procedure of $\mathcal{S}_{i,t}^{CQI}$ from the frequency domain to the CQI Mass Mode domain. The idea is to get some general information about the channel quality for the whole CQI report at each TTI *t*. From the viewpoint of the scheduler controller is enough to have statistics about the number of each CQI values averaged over $|\mathcal{B}|$ number of RBs. The principle of CQI-PS is exposed in Fig. C.1 and the mass value of the CQI report value is expressed in Eq. C.1:

$$MCQI_{i,v} = \frac{N_{CQI_v}}{|\mathcal{B}|}, \; CQI_v = 1,...,15 \tag{C.1}$$

where CQI_v represents the CQI quantized value as shown in Fig. B.12 from Appendix B. Unlike the mass mode, the normal mode CQI considers the frequency domain dimension in which the CQI value is normalized to its maximum value (e.g. 15 in LTE). The preprocessed CQI for Jakes and Vehicular A channel types are exposed in Fig. C.2.a and Fig. C.2.b by considering the same time range as shown in Figures B.12.a and B.12.b from Appendix B.

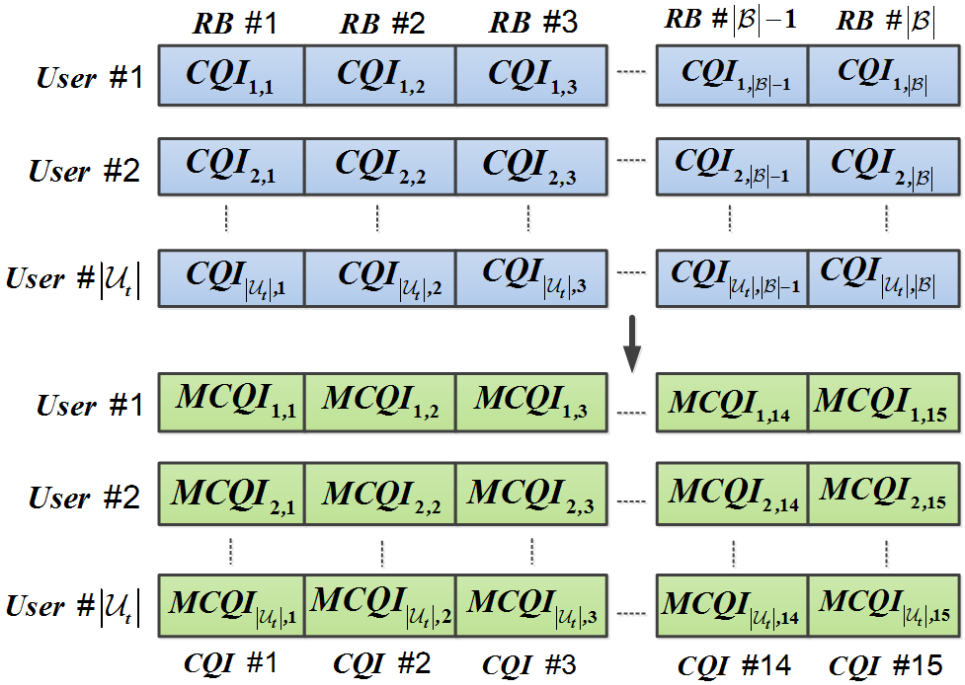
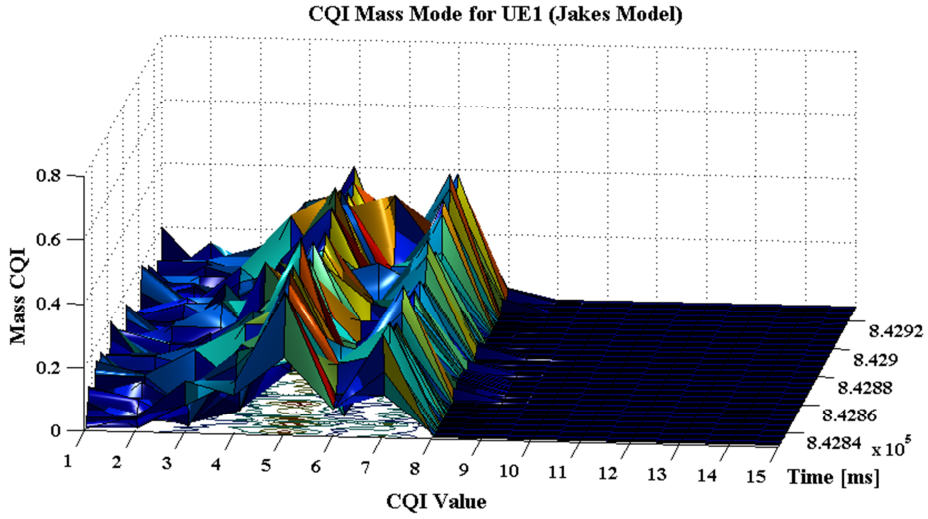
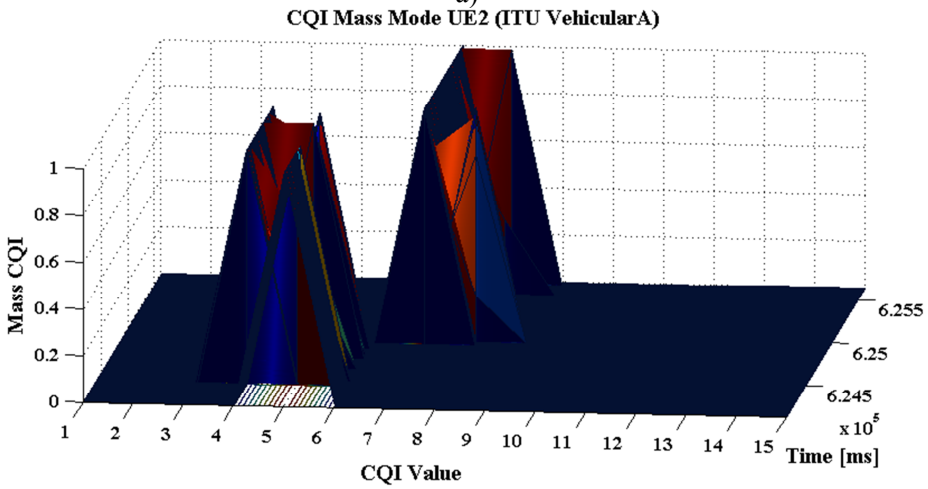


Fig. C.1 The Preprocessing Stage Principle



a)



b)

Fig. C.2 CQI Mass Mode for a) Jakes Model and b) ITU Vehicular A

The impact of the handover procedure can be sensed in Fig. C.2.b, where the CQI mass values increase due to the fact that UE2 starts to be served by the eNodeB with the highest SINR level. The Jakes fading model introduces severe fluctuations in the obtained CQI mass vector at each TTI. Let us define $\left|\mathcal{S}_{i,t}^{CQI}\right| = \left(N_{CQI}\right)^{|\mathcal{B}|}$ the initial CQI state space size of user $i \in \mathcal{U}_t$. After the preprocessing stage, the obtained state size of user $i \in \mathcal{U}_t$ is defined as indicated in Equation C.2:

$$\left|\mathcal{S}_{i,t}^{CQI,P}\right| = C_{|\mathcal{B}|+N_{CQI}-1}^{|\mathcal{B}|} \tag{C.2}$$

where $C_x^y = x!/(x-y)!\cdot y!$ represents the combination of x taken by y. The CQI-PS eliminates the duplicates in the CQI reports for each active user. For the

particular example of 1.4 MHz bandwidth $\left(N_{CQI}=15,\left|\mathcal{B}\right|=6\right)$, the initial CQI state space size is $\left|\mathcal{S}_{i,t}^{CQI}\right|=470.184.984.576$, and after the preprocessing stage, the CQI state space size becomes $\left|\mathcal{S}_{i,t}^{CQI,P}\right|=38.760$. Even for the lowest system bandwidth, after the CQI-PS stage, the number of possible combinations for the CQI states for one user is still very high.

The significant state space $\mathcal{S}_{i,t}^{CQI,P}$ size is due to the fact that low percentages of different CQI values are reported at each TTI especially in the Jakes fading model. For overhead reasons, this phenomenon can be avoided if and only if the top CQI values will be reported without a major degradation in the cell spectral efficiency. In this study, this process is executed at the eNodeB level after receiving the full CQI reports without any degradation of the system throughput (e.g. imperfect CQI reporting reasons). Therefore, only the significant percentage of the CQI mass values is considered by reducing much more the preprocessed CQI state $\mathcal{S}_{i,t}^{CQI,P}$ space size. Two methods are proposed to be analyzed in this sense: *Top Mass CQI* and *Majority Mass CQI* principles.

C.3 Top Mass CQI Principle

The idea aims to select the best percentages of CQI values for a given system bandwidth. Let us consider the top mass preprocessed CQI value set $\mathcal{S}_{i,t}^{CQI,P,vT}=\left\{MCQI_{vT,i}\left[t\right]\right\},\forall vT=1,..,N_{CQI}$ at TTI t. Then, the residual mass CQI set is: $\mathcal{S}_{i,t}^{CQI,P,ResT}=\mathcal{S}_{i,t}^{CQI,P}\setminus\bigcup_{vT}\mathcal{S}_{i,t}^{CQI,P,vT}$. When the top set $\mathcal{S}_{i,t}^{CQI,P,T}=\bigcup_{vT}\mathcal{S}_{i,t}^{CQI,P,vT}$ is

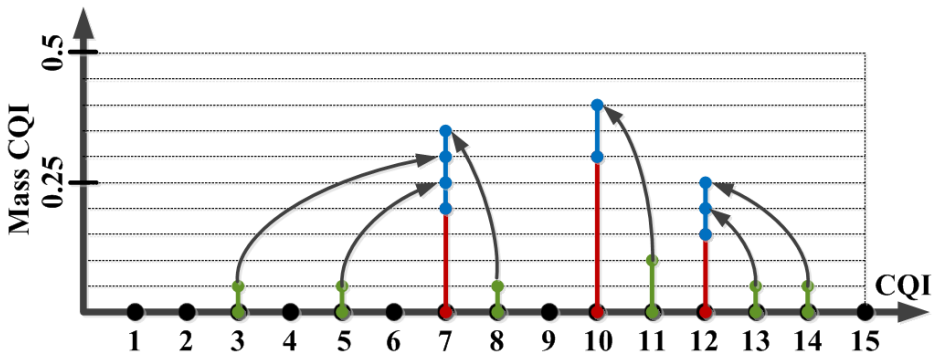


Fig. C.3 Minimum Distance based Top 3 Mass Mode CQI Reassignment

determined, the main problem is to add the percentages of mass CQIs from the residual set $\mathcal{S}_{i,t}^{CQI,P,ResT}$ to $\mathcal{S}_{i,t}^{CQI,P,T}$. This assignation process is entitled *minimum distance based mass CQI reassignment* which is presented in Fig. C.3. Based on the reassignment principle, the residual mass CQI values are associated to the nearest minimum top mass CQI value. The obtained state $\mathcal{S}_{i,t}^{CQI,P,RT}$ is entitled the *preprocessed CQI state space based on the reassigned top CQI mass mode*. The impact of the reassigned Top3 mass mode principle in the Jakes and Vehicular A fading models is depicted in Fig. C.4. It can be observed that, in the Jakes fast fading model, the amplitude of mass CQI increases to 0.6 when compared with the simple CQI preprocessing stage, by indicating that the percentage of neighboring CQI values are added to the top values. The result of the CQI reassigned Top 3 mass mode is more visible for the ITU Vehicular A fading in which the indices 6, 7, 8 are the most prevalent appearances in the CQI report.

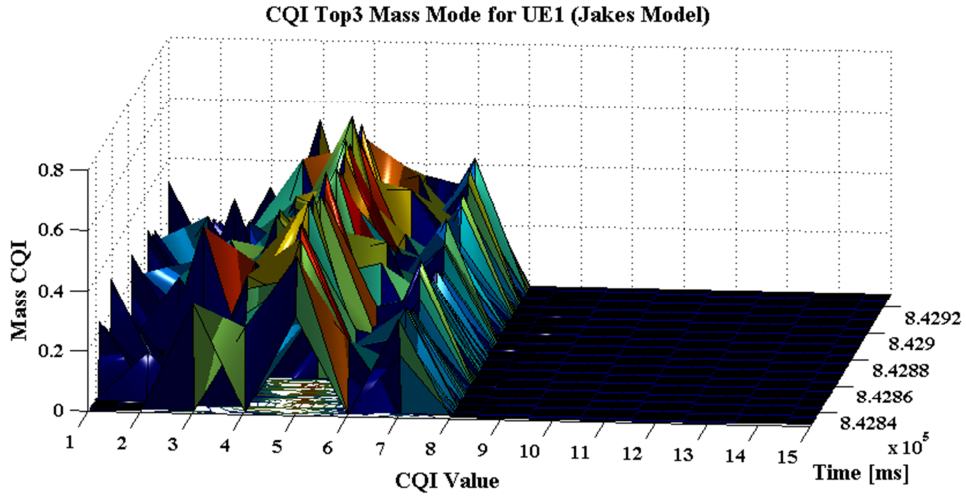
The main attention is focalized on the reassigned CQI state space size $\left|\mathcal{S}_{i,t}^{CQI,P,RT}\right|$ of the preprocessed CQI report. By following the principle from Eq. C.2, the generalized mathematical model for the space size of the preprocessed and the reassigned CQI state when the top CQI mass reassignment principle is used is denoted by Eq. C.3:

$$\left|\mathcal{S}_{i,t}^{CQI,P,RT}\right| = C_{|\mathcal{B}|+N_{CQI}-1}^{|\mathcal{B}|+\left(N_{CQI}-Top_{CQI}\right)} \cdot C_{N_{CQI}}^{Top_{CQI}} \tag{C.3}$$

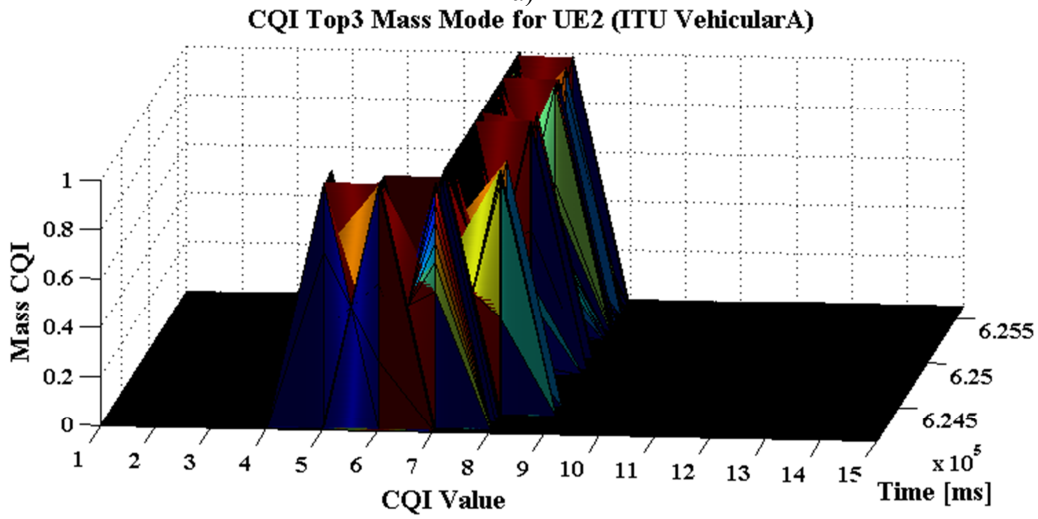
where Top_{CQI} is the required number of reassigned mass CQI values.

C.4 Majority Mass CQI Principle

This principle determines the top mass CQI values based on the maximum percentage of allowable RBs Ptg_{RB}. Precisely, if the sum of the best mass CQI values is greater than $\left(Ptg_{RB}+1\right)\%$, then the corresponding CQI indices form the majority class. Therefore, the reassignment procedure for the unselected CQI values is achieved by following the same principle as exposed in Fig. C.3. This method is a dynamic top mass CQI principle in the sense that the top of



a)



b)

Fig. C.4 CQI Top 3 Mass Mode for a) Jakes Model and b) ITU Vehicular A (Mode P:1-C:0-R:0)

the accepted CQI values varies at each TTI due to the stochastic nature of the fading model. If $\mathcal{S}_{i,t}^{CQI,P,vM} = \{MCQI_{vM}[t]\}, \forall vM = 1,..,N_{CQI}$ is the set of selected CQI values based on the majority sum mass CQI principle, vM is the CQI index corresponding to the selected mass CQI value and $\mathcal{S}_{i,t}^{CQI,P,M} = \bigcup_{vM} \mathcal{S}_{i,t}^{CQI,P,vM}$ is the preprocessed CQI majority mass state space, then the residual preprocessed CQI set becomes $\mathcal{S}_{i,t}^{CQI,P,ResM} = \mathcal{S}_{i,t}^{CQI,P} \setminus \mathcal{S}_{i,t}^{CQI,P,M}$ .Based on the reassignment procedure, the obtained CQI state space when performing the majority mass principle is $\mathcal{S}_{i,t}^{CQI,P,RM}$. The state $\mathcal{S}_{i,t}^{CQI,P,RM}$ space size can be calculated based on Eq. C.4:

$$\left|\mathcal{S}_{i,t}^{CQI,P,RM}\right| = C_{|\mathcal{B}|+N_{CQI}-1}^{|\mathcal{B}|+\left[N_{CQI}\cdot(1-Ptg_{RB})\right]} \cdot C_{N_{CQI}}^{\left[N_{CQI}\cdot(1-Ptg_{RB})\right]} \quad \text{(C.4)}$$

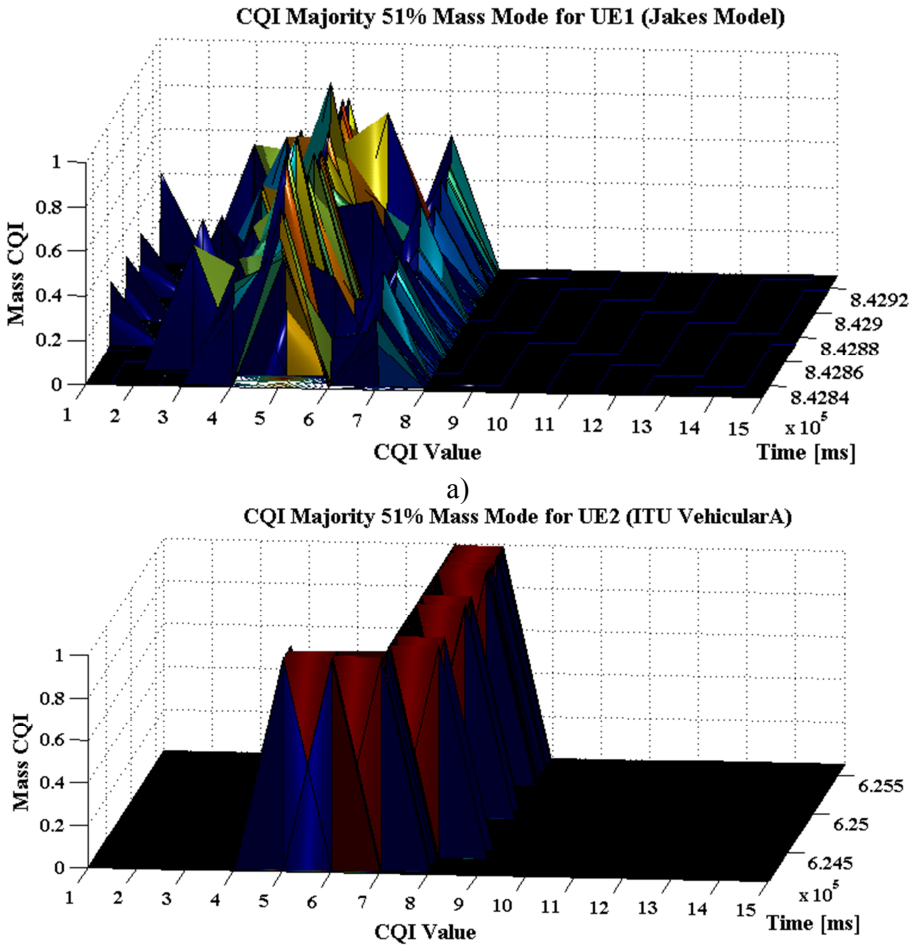


Fig. C.5 CQI Majority 51% Mass Mode for a) Jakes Model and b) ITU Vehicular A (Mode P:1-C:0-R:0)

where $\lfloor x \rfloor$ is the integer part of x. It is important to notice that in Chapter 4, it is used the notation of $\mathcal{S}_{i,t}^{CQI,P,TM}$ denoting the preprocessed CQI state based on top mass or majority mass reassignment principles. The impact of majority mass CQI principles for the Jakes and Vehicular Type A fading models when the maximum percentage is 51% is illustrated in Fig. C.5. The majority of 51% in the CQI mass mode impacts more visible in the Jakes model case when compared with the Top3 CQI mass mode denoting an amplitude of CQI mass value for the 5th CQI index of about 0.8. The same impact can be seen in the ITU Vehicular A model where the range of the selected mass CQI values is more restrictive. The algorithms for both top and majority mass modes based CQI-PS are analyzed by the Algorithm C.1.

As mentioned, the reassigned top or majority mass based CQI-PS aims to reduce the preprocessed CQI state space when compared with the previous

Algorithm C.1 Top Mass and Majority Mass in CQI State Space Classification

Requires:

$CT_{i,t}[v]$: vector that counts the number of apparitions of $v = 1,...,N_{CQI}$ in $\mathcal{S}_{i,t}^{CQI}$

$CT_{Top}[t]$: counts until reaches $Top_{CQI}[t]$

$CT_{Maj}[t]$: counts groups of RBs until reaches $(Ptg_{RB}+1)\%$

$Mask_{i,t}^{Top}[v]$: mask vector for $CT_{i,t}[v]$ based on $Top_{CQI}[t]$

$Mask_{i,t}^{Maj}[v]$: mask vector for $CT_{i,t}[v]$ based on $Ptg_{RB}[t]$

$maxCT_{i,T}$: the maximum number for a given CQI value

$d(v,v')$: distance between residual index and selected mass CQI index

for all active users $i = 1,...,|\mathcal{U}_t|$ **at each TTI** ***t***

$\mathcal{S}_{i,t}^{CQI} \leftarrow CQI\ State\ Space\ Module$

Update $CT_{i,t}[v] \leftarrow \mathcal{S}_{i,t}^{CQI}$

if (isTopMass) //Reassigned Top Mass Method

  while $CT_{Top}[t] \le Top_{CQI}[t]$

    for $v = 1,...,N_{CQI}$

      if $\left[\left(max\left(CT_{i,t}[v]\right) == true\right) \&\& \left(Mask_{i,t}^{Top}[v] \ne 1\right)\right]$

        $Mask_{i,t}^{Top}[v] = 1$

      end if

    end for

    $CT_{Top}[t]$++

  end while

else //Reassigned Majority Mass Method

  while $CT_{Maj}[t] \le Ptg_{RB}[t]$

    for $v = 1,...,N_{CQI}$

      if $\left[\left(max\left(CT_{i,t}[v]\right) == true\right) \&\& \left(Mask_{i,t}^{Maj}[v] \ne 1\right)\right]$

        $Mask_{i,t}^{Maj}[v] = 1$

        $maxCT_{i,t} = CT_{i,t}[v]$

      end if

    end for

    $CT_{Maj}[t] = CT_{Maj}[t] + maxCT_{i,t}$

  end while

end if

for $v, v' = 1,...,N_{CQI}$

  if $\left[(v \ne v') \&\& \left(min\left\{d(v,v')\right\}\right) \&\& \left(\left(Mask_{i,t}^{Top}[v] == 1\right) \| \left(Mask_{i,t}^{Maj}[v] == 1\right)\right)\right]$

    $MCQI_{i,t}^{RT}[v] = CT_{i,t}[v] + CT_{i,t}[v']$ //Add the residual $CT_{i,t}[v']$

    $MCQI_{i,t}^{RM}[v] = CT_{i,t}[v] + CT_{i,t}[v']$ //Add the residual $CT_{i,t}[v']$

  end if

end for

$\mathcal{S}_{i,t}^{CQI,P,RT} = MCQI_{i,vT}^{RT}[t] / |\mathcal{B}|$

$\mathcal{S}_{i,t}^{CQI,P,RM} = MCQI_{i,vP}^{RM}[t] / |\mathcal{B}|$

end for

Table C.1 CQI Top Mass Mode. Percentages of Different Schemes Reported to the $Top_{CQI}^{REF} = 15$ Case

Top CQI / Bandwidth	$Top_{CQI} = 9$	$Top_{CQI} = 8$	$Top_{CQI} = 7$	$Top_{CQI} = 6$	$Top_{CQI} = 5$	$Top_{CQI} = 4$	$Top_{CQI} = 3$	$Top_{CQI} = 2$	$Top_{CQI} = 1$
$N_{RB} = 6$	1626625	1287000	643500	200200	37537.5	4014.705882	223.0392157	5.417956656	0.03869969
$N_{RB} = 16$	20143.96	9008.48	2627.473	490.4617	56.59173509	3.810891252	0.136103259	0.0021661	1.03148E-05
$N_{RB} = 27$	1357.029	410.5297	82.10593	10.64336	0.86297529	0.041290684	0.001058735	1.22162E-05	4.25651E-08
$N_{RB} = 55$	27.06873	4.490665	0.498963	0.036383	0.001679202	4.6259E-05	6.90433E-07	4.6862E-09	9.70227E-12
$N_{RB} = 75$	4.631172	0.580914	0.048993	0.002722	9.60641E-05	2.03095E-06	2.33443E-08	1.22435E-10	1.96525E-13
$N_{RB} = 100$	0.881058	0.084695	0.005489	0.000235	6.40971E-06	1.04991E-07	9.37421E-10	3.82881E-12	4.79801E-15

Table C.2 CQI Majority Mass Mode. Percentages of Different Schemes Reported to the $Ptg_{RB}^{REF} = 0$ Case

Ptg. RBs / Bandwidth	$Ptg_{RB} = 0.9$	$Ptg_{RB} = 0.8$	$Ptg_{RB} = 0.7$	$Ptg_{RB} = 0.6$	$Ptg_{RB} = 0.5$	$Ptg_{RB} = 0.4$	$Ptg_{RB} = 0.3$	$Ptg_{RB} = 0.2$	$Ptg_{RB} = 0.1$
$N_{RB} = 6$	34125	197166.7	1301300	1626625	643500	200200	4014.705882	223.0392157	0.03869969
$N_{RB} = 16$	6245.098	17091.85	29544.48	20143.96	2627.473	490.4617	3.810891252	0.136103259	1.03148E-05
$N_{RB} = 27$	2353.448	4079.31	2985.463	1357.029	82.10593	10.64336	0.041290684	0.001058735	4.25651E-08
$N_{RB} = 55$	598.6842	536.7514	110.0795	27.06873	0.498963	0.036383	4.6259E-05	6.90433E-07	9.70227E-12
$N_{RB} = 75$	326.555	217.7033	25.00833	4.631172	0.048993	0.002722	2.03095E-06	2.33443E-08	1.96525E-13
$N_{RB} = 100$	185.498	93.64946	6.226145	0.881058	0.005489	0.000235	1.04991E-07	9.37421E-10	4.79801E-15

approach $\left(\left|\mathcal{S}_{i,t}^{CQI,P,TM}\right| < \left|\mathcal{S}_{i,t}^{CQI,P}\right|\right)$, where $\mathcal{S}_{i,t}^{CQI,P,TM}$ can be the reassigned top or majority CQI mass mode state space. But this reasoning is not always valid for all the possible inputs $\left(N_{CQI}, |\mathcal{B}|, Top_{CQI}, Ptg_{RB}\right)$ as indicated by Tables C.1 and C.2. Different combinations are highlighted in terms of $\left(\left|\mathcal{S}_{i,t}^{CQI,P,TM}\right| / \left|\mathcal{S}_{i,t}^{CQI,P}\right| \times 100\right)\%$. When the percentage is higher than 100% (red color), the proposed top or majority based CQI-PS does not bring any reduction in the original state space. This aspect is logical since the both methods introduce variability in the system bandwidth and thus, the space size may increase. The suggested operating input sets for the preprocessed CQI state space size reduction is depicted in green color by showing the percentage of how much the space size can be reduced when the preprocessing majority or top mass methods are used when compared with the initial preprocessed state space size.

The preprocessed CQI state space sizes $\left|\mathcal{S}_{i,t}^{CQI,P,RT}\right|$ and $\left|\mathcal{S}_{i,t}^{CQI,P,RM}\right|$ are pure theoretical since in practice, the upper bounds of Eq. C.3 and C.4 are never reached. This issue is due to the fading model in which, only a part of the CQI values is reported at each TTI. In other words, the values of CQI 1 and CQI 15 will not be reported in the same CQI feedback. Even under the Jakes fast fading model, in the exposed example, the range of the reported CQI belongs to [1, 8] for the worst case. It can be concluded that based on the fast fading model type, the size of the preprocessed reassigned CQI top mass mode state $\left|\mathcal{S}_{i,t}^{CQI,P,RT}\right|$ and the size of the preprocessed CQI reassigned majority mass mode state $\left|\mathcal{S}_{i,t}^{CQI,P,RM}\right|$ can be further reduced of about $C_4^{Top_{CQI}} / C_{N_{CQI}}^{Top_{CQI}}$ and $C_4^{\left[4\cdot(1-Ptg_{RB})\right]} / C_{N_{CQI}}^{\left[N_{CQI}\cdot(1-Ptg_{RB})\right]}$, respectively, for the ITU Vehicular A model, and of about $C_8^{Top_{CQI}} / C_{N_{CQI}}^{Top_{CQI}}$ and $C_8^{\left[8\cdot(1-Ptg_{RB})\right]} / C_{N_{CQI}}^{\left[N_{CQI}\cdot(1-Ptg_{RB})\right]}$, respectively, for the Jakes fast fading model. Under these approaches, the collection procedure of the preprocessed CQI states based on top or majority mass mode reassignment principles is much more simplified due to the reduced preprocessed CQI state space size which is involved in the collection algorithm. This principle is explained in the following section.

C.5 Preprocessed CQI Data Collection

For the classification purposes, the set of preprocessed CQI data points has to be collected. Of course, the collection of all possibilities for top or majority based CQI mass mode values is in general impossible to be achieved. In this sense, Algorithm C.2 is developed as a special offline algorithm which is able to collect the preprocessed CQI mass data until it reaches a predefined termination condition. From the viewpoint of the simulation test case, a large number of users with different mobility models, speeds and fading models can be used in order to increase the probability of increasing the preprocessed CQI mass state space size $\left(\mathcal{S}_t^{CQI,P,TM}\left[N_{CQI}\right]\left[\left|\mathcal{S}_t^{CQI,P,TM}\right|\right]\right)$ as fast as possible. The question that remains in this

Algorithm C.2 Preprocessed CQI Data Collection

Require:

$\mathcal{S}_t^{CQI,P,TM}\left[N_{CQI}\right]\left[\left|\mathcal{S}_t^{CQI,P,TM}\right|\right]$: stores all possible combinations of $\mathcal{S}_t^{CQI,P,TM}$

$TS_t^{CQI,P,TM}\left[N_{CQI}\right]\left[\left|\mathcal{S}_t^{CQI,P,TM}\right|\right]$: stores all possible combinations of $\mathcal{S}_t^{CQI,P,TM}$ and new CQIs

$Match\left(\mathcal{S}_{i,t}^{CQI,P,TM}\right)$: detects the matching columns between $\mathcal{S}_t^{CQI,P,TM}$ and $TS_t^{CQI,P,TM}$

1. for all active users $i=1,...,\left|\mathcal{U}_t\right|$ **at each TTI** ***t***
2. **for** $v=1,...,N_{CQI}$ *and* $s=1,...,\left|\mathcal{S}_t^{CQI,P,TM}\right|$
3. $TS_t^{CQI,P,TM}[v][s]=\mathcal{S}_{i,t}^{CQI,P,TM}[v]$
4. **end for**
5. end for
6. for $s=1,...,\left|\mathcal{S}_t^{CQI,P,TM}\right|$ // eliminate duplicates
7. **for** $\left[\left(s'<s\right)\&\&\left(Match\left(\mathcal{S}_{i,t}^{CQI,P,TM}\right)=false\right)\right]$
8. **for** $v=1,...,N_{CQI}$
9. **if** $\left(TS_t^{CQI,P,TM}[v][s']==TS_t^{CQI,P,TM}[v][s]\right)$
10. $Match\left(\mathcal{S}_{i,t}^{CQI,P,TM}\right)=true$
11. **end if**
12. **end for**
13. **end for**
14. **if** $\left(Match\left(\mathcal{S}_{i,t}^{CQI,P,TM}\right)=false\right)$
15. **for** $v=1,...,N_{CQI}$
16. $\mathcal{S}_t^{CQI,P,TM}[v][s'']=TS_t^{CQI,P,TM}[v][s]$
17. **end for**
18. $s''=s''+1$
19. **end if**
20. end for

case is regarded to the period when the algorithm should stop the collection procedure. As mentioned, a special termination condition of the preprocessed CQI data collection is proposed. The main principle is to count how many new $\mathcal{S}_{i,t}^{CQI,P,TM}$ observations are detected in comparison with the previous TTI. The counter value is averaged over the exponential filter as indicated by Eq. C.5:

$$\left|\overline{\mathcal{S}_{New,t}^{CQI,P,TM}}\right| = (1-\beta_{TM})\cdot\left|\overline{\mathcal{S}_{New,t-1}^{CQI,P,TM}}\right| + \beta_{TM}\cdot\left[|\mathcal{U}_t| - \left(\left|\mathcal{S}_{New,t}^{CQI,P,TM}\right| - \left|\mathcal{S}_{New,t-1}^{CQI,P,TM}\right|\right)/|\mathcal{U}_t|\right] \quad \text{(C.5)}$$

where $\left|\mathcal{S}_{New,t}^{CQI,P,TM}\right|$ represents the number of distinct elements which can be detected at each TTI t and $\left|\overline{\mathcal{S}_{New,t}^{CQI,P,TM}}\right|$ denotes its average value at each TTI t. When $\left|\overline{\mathcal{S}_{New,t}^{CQI,P,TM}}\right|$ exceeds a given threshold, let us say 0.99, then the entire collection algorithm is interrupted at TTI t.

C.6 Summary

The dimension of the CQI state for each user depends on the system bandwidth. Also, the CQI state space size is very high and it is impossible to be used in order to approximate optimal scheduling rules when the reinforcement learning principle is used. In this sense, the preprocessing stage is applied to the CQI state space in order to reduce the dimension and the size of the input spate space. Due to the characteristics of the fading processes, the size of the preprocessed CQI state space can be furthermore reduced by applying innovative reassignment procedures such as top mass and majority mass methods. The preprocessed observations are stored by using an innovative collection algorithm. The obtained collections for each bandwidth are used for the clustering algorithms and serve as validations sets for the RBFNN training procedure. Therefore, the preprocessed CQI observations are classified under different patterns and the obtained classified state space depends only on the number of preprocessed CQI centers being obtained based on the k-means clustering approaches.

Appendix D

Performance Evaluation of Clustering Algorithms for Different Bandwidths

D.1 Appendix Outline

The preprocessed CQI observations are collected for each LTE bandwidth and the best sets of CQI data centers are determined by using the clustering algorithms. The optimal set of centers is reached when the mean distortion from each collected data point to the obtained centers reaches the minimum value. This section analyzes the advantages of using the SAST based k-means clustering when compared with the classical heuristic approaches such as: Lloyd, Swap, Hybrid-EZ and Hybrid-SA based k-means clustering. The simulation environment keeps the same parameters as indicated in Section 4.7 from Chapter 4. The results are labeled for static numbers of centers. The best average distortion is monitored among the stage numbers. Also, the computational complexity is measured at the end of each simulation. The heuristic approach which obtains the best set of CQI centers by minimizing the best average distortion at each stage is considered to be the best option and it is marked in green colour. The best distortion and the best complexity for each algorithm are reported to the Lloyd approach by indicating any decrease or increase of the performance parameters at each monitored stage.

D.2 K-Means Clustering for 1.4 MHz

TABLE D.1 CQI Top Mass Mode based K-Means Clustering for 1.4 MHz

CQI Set	N_{CT}	Method	Best Average Distortion					CPU Time
			Stage 100	Stage 250	Stage 500	Stage 750	Stage 1000	
Top 3 2373	64	Lloyd	0.068573	0.068573	0.068573	0.068573	0.068573	8.31
		Swap	58.79%	49.91%	46.60%	43.89%	42.21%	-7.22%
		HEZ	-0.81%	-0.81%	-0.81%	-0.81%	-0.81%	-1.20%
		HSA	-0.99%	-4.30%	-4.89%	-5.06%	-5.98%	-3.49%
		HSAST	-0.72%	-5.93%	-6.37%	-7.87%	-7.94%	-3.49%
	128	Lloyd	0.0494672	0.0494672	0.0482168	0.0482168	0.0482168	15.97
		Swap	66.59%	58.12%	53.58%	49.48%	47.82%	-11.96%
		HEZ	0.70%	0.70%	2.95%	2.95%	2.95%	-0.44%
		HSA	-1.93%	-4.06%	-3.30%	-4.73%	-5.36%	-2.88%
		HSAST	-0.48%	-4.62%	-3.79%	-4.03%	-4.30%	-2.44%
	256	Lloyd	0.0366414	0.0366414	0.036184	0.036184	0.036184	31.32
		Swap	74.24%	65.42%	60.46%	57.39%	54.64%	-14.05%
		HEZ	0.79%	0.79%	2.07%	2.07%	2.07%	1.21%
		HSA	-0.02%	-1.47%	-1.39%	-2.28%	-2.73%	-2.59%
		HSAST	-0.02%	-1.24%	-0.67%	-1.13%	-1.43%	-0.96%
	512	Lloyd	0.0266967	0.0264421	0.0264421	0.0263782	0.0263782	61.04
		Swap	79.16%	77.92%	74.20%	71.12%	69.65%	-12.29%
		HEZ	-0.24%	0.72%	0.72%	0.96%	0.96%	-0.46%
		HSA	-0.46%	-1.35%	-2.77%	-4.58%	-5.86%	0.26%
		HSAST	-0.36%	-1.16%	-2.95%	-5.20%	-6.39%	-0.07%
	1024	Lloyd	0.0172527	0.0170946	0.017048	0.0169447	0.0168873	113.44
		Swap	86.92%	86.53%	85.94%	85.28%	84.87%	-9.27%
		HEZ	0.22%	1.15%	1.07%	1.69%	2.03%	1.18%
		HSA	0.25%	-1.15%	-1.50%	-1.84%	-2.80%	0.01%
		HSAST	-0.56%	-1.18%	-2.27%	-2.75%	-3.42%	0.97%
Top 4 5070	64	Lloyd	0.0612382	0.0612382	0.060589	0.060589	0.060589	18.13
		Swap	64.65%	58.90%	57.67%	55.44%	54.21%	5.74%
		HEZ	2.38%	1.99%	1.49%	1.49%	1.49%	-2.98%
		HSA	2.44%	0.02%	-1.27%	-1.51%	-1.61%	-7.23%
		HSAST	-0.69%	-1.05%	-0.34%	-0.55%	-0.70%	-7.34%
	128	Lloyd	0.0493443	0.0493443	0.0493443	0.049005	0.049005	34.45
		Swap	68.89%	62.63%	56.02%	54.57%	53.27%	0.38%
		HEZ	0.81%	0.81%	0.42%	1.12%	0.78%	-5.52%
		HSA	1.27%	0.13%	-1.18%	-1.30%	-1.56%	-9.99%
		HSAST	0.27%	-1.25%	-1.88%	-1.89%	-2.25%	-9.81%
	256	Lloyd	0.0410393	0.0406452	0.0406395	0.040592	0.040592	67.94
		Swap	62.21%	58.39%	52.63%	50.08%	48.42%	-8.99%
		HEZ	-0.47%	0.50%	0.51%	0.63%	0.63%	-2.83%
		HSA	-0.28%	0.36%	0.09%	-0.24%	-0.69%	-2.93%
		HSAST	-0.34%	0.23%	-0.27%	-0.75%	-1.18%	-3.55%
	512	Lloyd	0.0317743	0.031637	0.031637	0.031637	0.031539	136.45
		Swap	69.53%	66.03%	62.67%	60.47%	59.07%	-15.22%
		HEZ	0.85%	1.29%	1.29%	1.29%	1.60%	-1.69%
		HSA	0.65%	0.63%	-0.33%	-1.68%	-1.56%	-2.05%
		HSAST	0.88%	0.63%	-0.60%	-0.99%	-1.30%	-1.35%
	1024	Lloyd	0.0243293	0.0243293	0.0243293	0.0241685	0.0241685	261.94
		Swap	81.28%	79.84%	77.45%	77.48%	75.73%	-17.98%
		HEZ	0.56%	0.14%	-0.02%	0.28%	0.28%	-0.02%
		HSA	0.23%	-0.81%	-1.87%	-2.16%	-2.85%	1.23%
		HSAST	0.30%	-0.52%	-1.64%	-2.03%	-2.62%	0.15%
	64	Lloyd	0.064155	0.0636204	0.0632954	0.0632954	0.0632954	37.44
		Swap	54.30%	49.87%	45.98%	43.19%	40.92%	-4.43%
		HEZ	1.04%	0.51%	0.27%	0.27%	0.27%	-2.75%
		HSA	0.49%	0.27%	-0.29%	-0.72%	-0.87%	-3.85%
		HSAST	0.49%	0.29%	-0.09%	-0.63%	-0.76%	-7.48%
		Lloyd	0.0523319	0.0523319	0.0520625	0.0520625	0.0520625	71.95
		Swap	57.98%	51.70%	47.58%	43.69%	41.87%	-10.48%

Top 5 7294	128	HEZ	1.13%	0.80%	0.70%	0.49%	0.49%	-3.06%
		HSA	1.35%	-0.17%	-0.17%	-0.58%	-0.95%	-4.99%
		HSAST	1.32%	-0.27%	-0.07%	-0.19%	-0.40%	-4.95%
	256	Lloyd	0.0438612	0.0438476	0.0438119	0.0438119	0.0438119	143.31
		Swap	54.69%	49.08%	44.95%	41.41%	39.23%	-18.04%
		HEZ	-0.18%	-0.32%	-0.28%	-0.28%	-0.28%	-1.12%
		HSA	-0.44%	-0.70%	-0.75%	-1.05%	-1.24%	-1.47%
		HSAST	-0.17%	-0.71%	-0.85%	-1.27%	-1.39%	-2.06%
	512	Lloyd	0.035606	0.0355977	0.0355977	0.0355977	0.0355977	280.94
		Swap	53.13%	50.02%	46.13%	43.72%	41.71%	-21.65%
		HEZ	-0.43%	-0.58%	-0.62%	-0.62%	-0.62%	-1.14%
		HSA	-0.19%	-0.59%	-1.40%	-1.98%	-2.23%	-1.45%
		HSAST	-0.19%	-0.63%	-1.17%	-1.40%	-1.82%	-1.73%
	1024	Lloyd	0.0269966	0.0268453	0.0268453	0.0268453	0.0268453	542.87
		Swap	67.59%	65.76%	62.96%	60.83%	59.15%	-21.32%
		HEZ	-0.22%	0.18%	0.18%	0.18%	0.05%	0.72%
		HSA	-0.43%	-0.19%	-1.07%	-1.57%	-2.23%	1.41%
		HSAST	-0.29%	-0.59%	-1.68%	-2.54%	-2.77%	1.36%

D.3 K-Means Clustering for 3 MHz

TABLE D.2 CQI Top Mass Mode based K-Means Clustering for 3 MHz

CQI Set	N_{CT}	Method	Best Average Distortion					CPU Time
			Stage 100	Stage 250	Stage 500	Stage 750	Stage 1000	
Top 3 8749	64	Lloyd	0.0546572	0.0545916	0.0545397	0.0545397	0.0543819	20.58
		Swap	58.81%	47.54%	40.79%	38.65%	37.64%	-1.85%
		HEZ	2.71%	0.86%	0.95%	0.29%	0.58%	1.07%
		HSA	1.55%	-0.53%	-3.15%	-3.60%	-4.31%	-3.26%
		HSAST	-1.52%	-3.39%	-4.05%	-4.47%	-4.47%	-3.84%
	128	Lloyd	0.0394045	0.039241	0.0391172	0.0391172	0.0391172	36.45
		Swap	70.40%	55.12%	47.66%	40.19%	36.79%	-1.87%
		HEZ	2.43%	2.86%	3.18%	3.18%	3.18%	2.77%
		HSA	0.01%	-0.98%	-1.59%	-1.87%	-2.70%	-0.66%
		HSAST	0.14%	-1.07%	-2.57%	-3.25%	-3.62%	-2.77%
	256	Lloyd	0.0290246	0.0286324	0.028595	0.0285145	0.0285145	67.88
		Swap	59.59%	52.18%	42.66%	37.65%	34.61%	-6.73%
		HEZ	0.66%	1.01%	1.14%	1.42%	0.51%	0.81%
		HSA	0.32%	0.00%	-2.56%	-3.66%	-4.74%	2.09%
		HSAST	-0.80%	-1.57%	-4.52%	-5.49%	-6.40%	2.50%
	512	Lloyd	0.0187032	0.0187032	0.0183847	0.0183847	0.0183847	126.03
		Swap	72.89%	63.44%	55.28%	49.13%	44.67%	-7.46%
		HEZ	0.08%	0.08%	1.68%	0.81%	0.81%	1.41%
		HSA	-0.97%	-2.87%	-3.35%	-5.52%	-6.53%	2.07%
		HSAST	-1.50%	-3.29%	-3.78%	-5.35%	-5.95%	2.18%
	1024	Lloyd	0.0111749	0.0110497	0.0110346	0.0110346	0.0110346	232.63
		Swap	82.75%	76.13%	66.53%	59.37%	53.90%	-7.36%
		HEZ	0.32%	0.25%	-0.30%	-0.30%	-0.30%	1.04%
		HSA	0.09%	0.11%	-2.91%	-4.24%	-5.00%	0.91%
		HSAST	-0.14%	-0.56%	-2.98%	-4.58%	-5.49%	1.10%
	64	Lloyd	0.0451887	0.0451887	0.0449437	0.0448272	0.0448272	68.52
		Swap	46.95%	40.10%	38.67%	36.60%	36.26%	8.06%
		HEZ	0.56%	-0.30%	0.25%	0.51%	0.51%	-2.92%
		HSA	-0.49%	-1.96%	-1.74%	-1.72%	-1.72%	-7.63%
		HSAST	0.40%	-1.55%	-1.51%	-1.77%	-1.87%	-6.68%
	128	Lloyd	0.033872	0.0338571	0.0338571	0.0338571	0.0338571	114.68
		Swap	57.05%	51.13%	41.04%	38.06%	35.89%	5.96%
		HEZ	1.40%	1.44%	0.94%	0.94%	0.53%	-1.58%
		HSA	1.45%	0.04%	-0.65%	-1.23%	-1.61%	-3.98%
		HSAST	1.05%	-0.61%	-1.34%	-1.54%	-1.80%	-3.92%

Top 4 38685	256	Lloyd	0.0259889	0.0259889	0.025883	0.025883	0.025883	200.2
		Swap	60.79%	50.85%	42.81%	38.56%	36.23%	3.85%
		HEZ	0.70%	0.49%	-0.07%	-0.07%	-0.07%	0.70%
		HSA	0.12%	-0.29%	-0.76%	-1.64%	-2.07%	-3.26%
		HSAST	-0.43%	-1.12%	-1.61%	-1.88%	-2.31%	-3.15%
	512	Lloyd	0.0197449	0.0196595	0.0196595	0.0196595	0.0196595	367.29
		Swap	58.15%	53.27%	45.90%	42.22%	39.63%	-1.93%
		HEZ	0.01%	0.32%	0.32%	0.32%	-0.02%	0.71%
		HSA	0.34%	-0.30%	-0.88%	-1.50%	-1.96%	-0.86%
		HSAST	-0.40%	-0.61%	-1.42%	-1.69%	-1.80%	-0.27%
	1024	Lloyd	0.0147076	0.0146225	0.0146225	0.0146225	0.0146183	697.39
		Swap	60.01%	57.81%	53.23%	49.90%	46.99%	-5.62%
		HEZ	-0.36%	0.22%	0.16%	0.16%	0.19%	0.87%
		HSA	-0.49%	-0.58%	-1.03%	-1.43%	-1.89%	0.40%
		HSAST	-0.57%	-0.76%	-1.32%	-1.90%	-2.15%	-0.02%
Top 5 74433	64	Lloyd	0.0354572	0.0352753	0.0351903	0.0351903	0.03517	121.46
		Swap	50.83%	42.29%	37.01%	34.31%	33.89%	17.17%
		HEZ	1.86%	1.93%	1.04%	1.04%	1.10%	-2.97%
		HSA	1.10%	0.29%	-0.21%	-0.21%	-0.25%	-6.48%
		HSAST	0.52%	0.43%	-0.15%	-0.68%	-0.89%	-7.91%
	128	Lloyd	0.0279971	0.0278075	0.0276578	0.0276578	0.0276578	195.98
		Swap	49.95%	44.21%	39.43%	37.70%	36.30%	14.01%
		HEZ	0.56%	0.99%	1.11%	1.08%	1.08%	-4.49%
		HSA	0.17%	-0.23%	-0.06%	-0.34%	-0.46%	-5.91%
		HSAST	0.10%	-0.27%	0.05%	-0.14%	-0.29%	-7.63%
	256	Lloyd	0.0220924	0.0220578	0.0219622	0.0219622	0.0219622	335.32
		Swap	50.72%	46.46%	42.70%	39.55%	36.94%	10.14%
		HEZ	-0.08%	-0.25%	0.19%	0.19%	0.19%	-3.55%
		HSA	-0.36%	-0.41%	-0.49%	-0.86%	-1.24%	-5.39%
		HSAST	-0.76%	-1.09%	-1.02%	-1.20%	-1.24%	-4.78%
	512	Lloyd	0.0174605	0.0174343	0.0173937	0.0173916	0.0173916	601.66
		Swap	52.09%	48.74%	44.01%	41.78%	40.15%	6.54%
		HEZ	-0.43%	-0.42%	-0.27%	-0.33%	-0.33%	-2.18%
		HSA	-0.28%	-0.43%	-0.61%	-0.92%	-1.10%	-2.95%
		HSAST	-0.56%	-0.84%	-1.14%	-1.40%	-1.59%	-3.73%
	1024	Lloyd	0.0136995	0.0136995	0.0136995	0.0136995	0.0136839	1130.28
		Swap	54.27%	51.69%	48.77%	46.52%	44.69%	1.98%
		HEZ	0.06%	-0.09%	-0.12%	-0.88%	-0.77%	-2.51%
		HSA	0.12%	-0.27%	-1.09%	-1.39%	-1.49%	-3.18%
		HSAST	-0.01%	-0.54%	-1.18%	-1.53%	-1.62%	-2.87%

D.4 K-Means Clustering for 5 MHz

TABLE D.3 CQI Top Mass Mode based K-Means Clustering for 5 MHz

CQI Set	N_{CT}	Method	Best Average Distortion					CPU Time
			Stage 100	Stage 250	Stage 500	Stage 750	Stage 1000	
Top 3 13016	64	Lloyd	0.0493619	0.0488598	0.0488243	0.0488243	0.0488243	23.71
		Swap	51.74%	45.51%	36.10%	34.56%	32.70%	-0.21%
		HEZ	1.52%	2.57%	2.33%	2.33%	2.33%	-0.89%
		HSA	-0.76%	0.00%	-3.38%	-3.40%	-3.40%	-3.71%
		HSAST	-1.67%	-1.59%	-3.20%	-3.28%	-3.84%	-1.01%
	128	Lloyd	0.0351982	0.0341397	0.0341397	0.0341397	0.0341397	40.76
		Swap	56.57%	49.48%	42.02%	35.51%	32.10%	-2.40%
		HEZ	0.01%	3.11%	3.11%	3.11%	3.11%	-1.08%
		HSA	1.03%	2.35%	-0.20%	-1.08%	-2.28%	0.42%
		HSAST	1.11%	-0.18%	-1.69%	-1.86%	-2.33%	0.56%
		Lloyd	0.0217362	0.0217362	0.0217362	0.0217362	0.0217362	71.92
		Swap	76.95%	58.95%	48.06%	39.29%	35.57%	-5.19%

	256	HEZ	3.17%	0.63%	-0.38%	-0.38%	-0.38%	0.57%
		HSA	3.69%	-1.26%	-3.92%	-4.95%	-5.33%	0.13%
		HSAST	3.70%	0.43%	-2.37%	-4.37%	-4.87%	-0.43%
	512	Lloyd	0.0129848	0.0128822	0.0128747	0.012858	0.012858	126.86
		Swap	89.85%	73.00%	58.68%	50.09%	43.34%	-4.70%
		HEZ	2.00%	2.57%	0.51%	0.58%	0.58%	0.06%
		HSA	1.42%	-0.31%	-3.23%	-4.58%	-5.76%	2.22%
		HSAST	0.66%	-0.84%	-3.02%	-4.10%	-5.20%	0.13%
	1024	Lloyd	0.0072355	0.00716648	0.00716648	0.00716648	0.00716648	228.81
		Swap	94.85%	87.31%	73.36%	63.30%	55.50%	-2.46%
		HEZ	1.77%	2.75%	2.70%	1.38%	0.36%	2.16%
		HSA	1.55%	1.02%	-1.10%	-2.45%	-4.22%	0.39%
		HSAST	1.92%	0.31%	-1.39%	-2.56%	-4.08%	1.28%
Top 4 65691	64	Lloyd	0.036039	0.0359828	0.035865	0.0357592	0.0356395	89.38
		Swap	48.06%	39.76%	35.73%	33.78%	32.71%	10.80%
		HEZ	1.00%	1.04%	1.21%	1.51%	1.84%	-1.13%
		HSA	3.58%	2.61%	0.73%	0.23%	-0.69%	-3.96%
		HSAST	2.40%	-0.16%	-0.29%	-0.48%	-0.69%	-3.42%
	128	Lloyd	0.0267721	0.026764	0.026764	0.026764	0.0267461	139.41
		Swap	53.70%	42.34%	37.31%	34.07%	32.76%	5.23%
		HEZ	1.29%	0.30%	0.30%	0.30%	0.37%	-2.62%
		HSA	0.83%	0.00%	-0.79%	-1.67%	-2.19%	-4.48%
		HSAST	-0.34%	-1.19%	-1.75%	-2.04%	-2.17%	-4.65%
	256	Lloyd	0.0198441	0.0198441	0.0197216	0.0197216	0.0197216	229.8
		Swap	54.01%	46.35%	40.72%	36.41%	34.09%	2.30%
		HEZ	0.19%	-0.62%	0.00%	-0.45%	-0.48%	-1.23%
		HSA	-0.07%	-1.62%	-1.84%	-2.32%	-2.46%	-1.31%
		HSAST	-0.91%	-1.66%	-1.73%	-2.34%	-2.46%	-2.53%
	512	Lloyd	0.0143951	0.0143291	0.0143291	0.0143291	0.0143291	402.4
		Swap	57.62%	50.74%	43.98%	39.34%	36.18%	-3.38%
		HEZ	-0.20%	-0.27%	-0.43%	-0.71%	-0.71%	-1.16%
		HSA	-0.05%	-0.81%	-1.89%	-2.59%	-2.83%	-0.60%
		HSAST	-0.64%	-1.16%	-2.11%	-2.79%	-3.03%	-1.46%
	1024	Lloyd	0.0102306	0.0101994	0.0101994	0.0101994	0.0101855	729.57
		Swap	56.35%	53.35%	47.98%	44.43%	41.30%	-6.29%
		HEZ	-0.94%	-0.88%	-0.98%	-0.98%	-0.84%	-0.81%
		HSA	-0.79%	-1.06%	-1.78%	-2.33%	-2.79%	-0.50%
		HSAST	-0.84%	-1.18%	-2.22%	-2.63%	-3.08%	-0.77%
Top 5 117214	64	Lloyd	0.0416194	0.0405143	0.0381164	0.0381164	0.0381164	203.14
		Swap	-11.36%	-11.77%	-9.57%	-13.37%	-13.97%	-7.47%
		HEZ	-37.02%	-35.92%	-31.88%	-31.88%	-31.93%	-19.14%
		HSA	-37.40%	-36.42%	-32.58%	-32.90%	-33.00%	-21.10%
		HSAST	-38.24%	-36.76%	-33.22%	-33.22%	-33.22%	-21.23%
	128	Lloyd	0.0201284	0.0199907	0.0199907	0.0199907	0.0199907	261.14
		Swap	43.44%	37.08%	33.40%	30.12%	28.74%	8.79%
		HEZ	0.44%	0.35%	0.02%	0.02%	0.02%	-2.39%
		HSA	0.61%	0.21%	-0.48%	-0.66%	-0.91%	-3.85%
		HSAST	-0.18%	-0.31%	-0.75%	-0.85%	-0.97%	-3.94%
	256	Lloyd	0.0156037	0.0156037	0.0156037	0.015561	0.0155517	429.9
		Swap	45.89%	39.73%	34.60%	31.98%	30.27%	4.83%
		HEZ	0.07%	-0.12%	-0.69%	-0.41%	-0.35%	-2.24%
		HSA	0.17%	-0.78%	-1.30%	-1.38%	-1.50%	-5.12%
		HSAST	-0.06%	-0.78%	-1.32%	-1.10%	-1.08%	-2.76%
	512	Lloyd	0.0121281	0.0121157	0.0120909	0.0120792	0.0120792	766.39
		Swap	45.66%	41.05%	36.92%	34.18%	31.59%	2.99%
		HEZ	-0.60%	-0.50%	-0.30%	-0.20%	-0.20%	-2.36%
		HSA	-0.23%	-0.84%	-1.08%	-1.29%	-1.43%	-3.90%
		HSAST	-0.25%	-0.90%	-1.29%	-1.36%	-1.51%	-4.22%
	1024	Lloyd	0.0093295	0.00930634	0.00930634	0.00930239	0.00929975	1362.03
		Swap	45.20%	42.19%	38.84%	36.43%	34.63%	-1.57%
		HEZ	-0.76%	-1.08%	-1.18%	-1.14%	-1.12%	-0.29%
		HSA	-0.28%	-1.16%	-1.47%	-1.83%	-2.00%	-1.35%
		HSAST	-0.32%	-1.04%	-1.59%	-1.83%	-2.06%	-1.35%

D.5 K-Means Clustering for 10 MHz

TABLE D.4 CQI Top Mass Mode based K-Means Clustering for 10 MHz

CQI Set	N_{CT}	Method	Best Average Distortion					CPU Time
			Stage 100	Stage 250	Stage 500	Stage 750	Stage 1000	
Top 3 18531	64	Lloyd	0.0381545	0.0379898	0.0371624	0.0371624	0.0371624	21.63
		Swap	49.49%	34.76%	31.74%	28.86%	28.03%	-1.90%
		HEZ	2.04%	1.69%	1.44%	1.44%	1.44%	-0.32%
		HSA	2.37%	-1.46%	-2.11%	-3.19%	-3.47%	-1.29%
		HSAST	0.10%	-4.28%	-3.25%	-3.62%	-3.72%	-0.42%
	128	Lloyd	0.022332	0.021893	0.0218627	0.0218627	0.0217414	35.15
		Swap	66.41%	46.96%	35.99%	29.36%	27.77%	-4.55%
		HEZ	-0.04%	-2.25%	-2.24%	-2.24%	-1.69%	-2.76%
		HSA	-2.28%	-2.24%	-4.65%	-6.66%	-6.46%	-2.19%
		HSAST	-2.25%	-3.12%	-5.11%	-6.51%	-7.47%	-1.99%
	256	Lloyd	0.0116992	0.0114778	0.0113658	0.0113658	0.0113658	56.9
		Swap	90.59%	67.31%	48.60%	40.66%	34.79%	-3.92%
		HEZ	6.84%	5.39%	4.92%	4.92%	3.96%	-1.18%
		HSA	3.50%	0.45%	-1.26%	-3.76%	-5.28%	-0.62%
		HSAST	2.82%	-1.20%	-4.57%	-5.87%	-6.23%	-0.44%
	512	Lloyd	0.0059073	0.00590733	0.00590733	0.00590733	0.00583646	98.58
		Swap	111.50%	87.79%	60.74%	48.80%	42.98%	-2.01%
		HEZ	1.05%	1.05%	1.05%	1.02%	1.76%	-1.26%
		HSA	-0.27%	-1.29%	-4.08%	-5.76%	-6.10%	-0.49%
		HSAST	-0.33%	-2.38%	-4.24%	-5.59%	-6.14%	-0.36%
	1024	Lloyd	0.0030405	0.00304052	0.00301569	0.0029993	0.0029993	179.61
		Swap	117.05%	105.38%	88.42%	72.62%	63.81%	-0.42%
		HEZ	-0.04%	-0.64%	-0.86%	-0.41%	-0.73%	-1.05%
		HSA	-0.97%	-1.69%	-2.92%	-4.23%	-4.92%	-1.05%
		HSAST	-0.77%	-3.35%	-4.20%	-5.47%	-6.04%	-0.70%
Top 4 90584	64	Lloyd	0.0244228	0.0244228	0.024397	0.0242147	0.0239813	85.08
		Swap	47.55%	37.01%	31.23%	28.80%	29.13%	1.54%
		HEZ	2.26%	2.24%	1.51%	1.38%	2.36%	-1.59%
		HSA	2.26%	-0.53%	-1.15%	-1.29%	-0.75%	-2.73%
		HSAST	0.13%	-1.48%	-2.56%	-1.97%	-1.01%	-3.50%
	128	Lloyd	0.0168714	0.0167634	0.0167226	0.0167226	0.0167226	126.24
		Swap	48.68%	36.80%	26.75%	25.44%	24.50%	-1.83%
		HEZ	1.29%	1.84%	0.60%	0.45%	0.16%	-2.12%
		HSA	2.63%	0.38%	-1.70%	-1.92%	-2.62%	-2.44%
		HSAST	-0.63%	-1.11%	-2.58%	-2.61%	-3.10%	-2.92%
	256	Lloyd	0.0113281	0.0112002	0.0110918	0.0110853	0.0110853	197.33
		Swap	54.36%	45.69%	37.77%	32.89%	30.66%	-1.84%
		HEZ	0.22%	-0.50%	0.47%	0.53%	0.18%	-1.61%
		HSA	-0.28%	-1.08%	-1.36%	-2.34%	-2.83%	-0.31%
		HSAST	-0.74%	-2.27%	-2.52%	-2.55%	-3.02%	1.60%
	512	Lloyd	0.0073305	0.00732301	0.00730473	0.00728902	0.00728902	326.88
		Swap	62.82%	50.39%	41.42%	37.04%	34.45%	-3.71%
		HEZ	0.84%	0.19%	-0.60%	-0.94%	-1.13%	-2.76%
		HSA	0.71%	-0.93%	-2.24%	-2.58%	-3.06%	-2.18%
		HSAST	-0.14%	-1.77%	-2.39%	-2.97%	-3.23%	0.06%
	1024	Lloyd	0.0047596	0.00475957	0.00475522	0.00473992	0.00470667	556.03
		Swap	69.23%	62.54%	54.39%	47.32%	43.48%	-3.12%
		HEZ	-1.05%	-1.35%	-1.26%	-0.98%	-0.70%	-1.11%
		HSA	-1.11%	-1.82%	-2.71%	-2.93%	-2.82%	-0.40%
		HSAST	-1.33%	-2.31%	-3.06%	-3.38%	-3.10%	-0.53%
	64	Lloyd	0.0168175	0.0165103	0.0164949	0.0164949	0.0164949	162.89
		Swap	37.18%	33.75%	27.36%	25.80%	25.29%	0.28%
		HEZ	-0.17%	0.55%	0.64%	0.64%	0.64%	-1.07%
		HSA	-1.20%	0.18%	-0.30%	-0.65%	-0.76%	-2.82%
		HSAST	-1.62%	-0.52%	-0.98%	-1.30%	-1.36%	-2.26%
		Lloyd	0.0123838	0.0123821	0.0123821	0.0123262	0.0123262	247.45
		Swap	44.36%	36.61%	26.27%	23.40%	21.91%	0.22%

Top 5 125470	128	HEZ	0.11%	-0.47%	-0.47%	-0.01%	-0.01%	-0.42%
		HSA	1.91%	-0.01%	-1.36%	-0.99%	-1.13%	-0.10%
		HSAST	0.28%	-0.47%	-1.45%	-1.21%	-1.32%	-0.04%
	256	Lloyd	0.0091659	0.00916585	0.00913622	0.00913622	0.00913622	400.22
		Swap	47.90%	34.64%	29.52%	26.44%	24.27%	-5.12%
		HEZ	0.13%	-0.61%	-0.28%	-0.62%	-0.62%	0.53%
		HSA	0.41%	-0.91%	-1.72%	-1.96%	-2.03%	-0.33%
		HSAST	0.38%	-1.21%	-1.51%	-1.58%	-1.75%	-0.50%
	512	Lloyd	0.0067629	0.00672949	0.00672949	0.00668693	0.00668693	674.17
		Swap	41.48%	37.71%	31.68%	29.48%	27.17%	-5.64%
		HEZ	-1.24%	-1.08%	-1.22%	-0.59%	-0.59%	0.25%
		HSA	-1.02%	-1.20%	-1.55%	-1.36%	-1.54%	0.44%
		HSAST	-0.98%	-1.23%	-1.53%	-1.33%	-1.49%	-0.50%
	1024	Lloyd	0.0048985	0.00489853	0.00489853	0.0048985	0.00489853	1148.29
		Swap	44.21%	40.45%	35.86%	33.15%	31.08%	-7.77%
		HEZ	-0.79%	-1.50%	-1.51%	-1.51%	-1.51%	0.36%
		HSA	-0.66%	-1.41%	-1.78%	-1.98%	-2.05%	-0.36%
		HSAST	-0.59%	-1.41%	-1.90%	-2.37%	-2.59%	-0.04%

D.6 K-Means Clustering for 15 MHz

TABLE D.5 CQI Top Mass Mode based K-Means Clustering for 15 MHz

CQI Set	N_{CT}	Method	Best Average Distortion					CPU Time
			Stage 100	Stage 250	Stage 500	Stage 750	Stage 1000	
Top 3 24268	64	Lloyd	0.0278604	0.0269703	0.0269205	0.0267886	0.0267886	18.33
		Swap	56.27%	40.32%	34.41%	33.77%	30.04%	-1.69%
		HEZ	0.07%	3.13%	3.32%	1.75%	1.75%	-0.55%
		HSA	0.00%	0.18%	-1.03%	-2.79%	-2.97%	-1.20%
		HSAST	-3.06%	-2.44%	-3.67%	-3.31%	-3.40%	-1.85%
	128	Lloyd	0.0144288	0.0143167	0.0141822	0.0141822	0.0139663	28.92
		Swap	72.32%	52.40%	35.54%	31.03%	26.22%	-2.07%
		HEZ	-1.07%	-0.29%	0.43%	-0.03%	-0.21%	-1.14%
		HSA	4.26%	-3.64%	-6.75%	-7.69%	-6.97%	-0.83%
		HSAST	0.02%	-4.67%	-7.28%	-8.48%	-7.71%	-1.80%
	256	Lloyd	0.0071712	0.00717116	0.00691605	0.00691605	0.00691605	47.32
		Swap	115.57%	75.38%	56.72%	44.61%	31.32%	-1.12%
		HEZ	4.50%	2.49%	6.27%	3.61%	1.82%	-1.08%
		HSA	1.07%	-4.25%	-3.65%	-4.31%	-6.12%	-0.68%
		HSAST	-0.37%	-3.87%	-4.31%	-5.91%	-6.98%	-0.46%
	512	Lloyd	0.0036172	0.00355709	0.00355709	0.00355709	0.00351961	82.85
		Swap	113.59%	96.58%	72.60%	55.53%	41.87%	-0.98%
		HEZ	-0.94%	0.44%	-1.02%	-1.02%	-0.36%	-1.68%
		HSA	-0.92%	-2.54%	-5.00%	-6.17%	-6.13%	-0.64%
		HSAST	-1.72%	-1.72%	-3.63%	-5.37%	-6.10%	-1.45%
	1024	Lloyd	0.0018732	0.00180647	0.00180271	0.00180271	0.00179314	149.42
		Swap	126.87%	125.17%	90.73%	90.73%	68.48%	0.12%
		HEZ	1.34%	3.24%	2.34%	1.37%	1.63%	-1.56%
		HSA	-0.70%	0.22%	-2.28%	-3.90%	-5.00%	-0.38%
		HSAST	-0.66%	0.03%	-2.23%	-3.18%	-3.84%	-0.96%
	64	Lloyd	0.0167841	0.0167841	0.0166911	0.0165301	0.0165301	67.28
		Swap	64.24%	34.62%	31.00%	27.39%	25.79%	1.63%
		HEZ	4.47%	1.93%	2.50%	3.49%	3.49%	-1.07%
		HSA	2.24%	-3.32%	-4.18%	-3.25%	-3.60%	-3.67%
		HSAST	0.30%	-2.46%	-4.20%	-3.47%	-3.58%	-3.02%
	128	Lloyd	0.0104763	0.0102228	0.0102228	0.0101931	0.0101931	103
		Swap	64.12%	47.01%	38.56%	32.97%	30.01%	0.62%
		HEZ	2.50%	1.38%	1.38%	1.17%	1.17%	-3.32%
		HSA	3.27%	2.01%	0.32%	-1.32%	-2.27%	-4.20%

Top 4 105755		HSAST	1.16%	1.20%	0.21%	-0.08%	-0.56%	-1.20%
	256	Lloyd	0.0065817	0.00657525	0.00654976	0.0065412	0.00653147	161.53
		Swap	67.74%	49.07%	35.84%	31.05%	27.43%	-2.77%
		HEZ	1.71%	0.44%	-0.76%	-0.91%	-0.76%	-2.47%
		HSA	2.83%	0.87%	-2.07%	-2.73%	-2.98%	-1.73%
		HSAST	1.26%	-0.90%	-2.63%	-3.04%	-3.36%	-0.63%
	512	Lloyd	0.0042498	0.00421634	0.00421634	0.00421634	0.00418854	258.52
		Swap	74.68%	64.60%	53.59%	41.43%	36.91%	-1.86%
		HEZ	-0.15%	0.10%	-0.33%	-0.33%	0.33%	-1.46%
		HSA	-0.33%	-0.51%	-2.36%	-3.27%	-3.03%	-1.12%
		HSAST	-0.99%	-2.78%	-3.25%	-3.90%	-3.53%	-0.41%
	1024	Lloyd	0.0027204	0.00272043	0.00272043	0.00271469	0.00271469	428.19
		Swap	66.25%	60.93%	51.87%	45.34%	42.17%	-1.36%
		HEZ	-0.33%	-0.33%	-0.33%	-0.12%	-0.13%	-1.43%
		HSA	0.10%	-0.31%	-1.82%	-2.43%	-3.04%	-0.59%
		HSAST	-0.01%	-0.84%	-1.90%	-2.53%	-2.97%	-0.06%
Top 5 162936	64	Lloyd	0.0119565	0.0119565	0.0119565	0.0119209	0.0119209	151.87
		Swap	45.51%	34.42%	27.80%	26.02%	23.36%	0.75%
		HEZ	3.85%	0.51%	0.51%	0.81%	0.81%	-2.82%
		HSA	3.60%	2.02%	-0.74%	-1.03%	-1.03%	-4.42%
		HSAST	0.86%	-0.86%	-1.91%	-2.09%	-2.09%	-5.02%
	128	Lloyd	0.0085816	0.00858156	0.00857429	0.00852074	0.00850757	229.39
		Swap	52.31%	39.85%	29.28%	26.90%	24.26%	-1.29%
		HEZ	-0.62%	-1.14%	-1.05%	-0.43%	-0.28%	-2.25%
		HSA	1.22%	-0.94%	-1.64%	-1.43%	-1.67%	-2.51%
		HSAST	0.37%	-1.63%	-2.14%	-1.79%	-1.92%	-2.91%
	256	Lloyd	0.0061749	0.00614828	0.00614828	0.0061442	0.0061442	355.65
		Swap	48.72%	38.48%	31.65%	28.16%	25.55%	-3.89%
		HEZ	0.65%	0.40%	-0.71%	-0.65%	-0.65%	-2.20%
		HSA	0.39%	-0.34%	-1.15%	-2.07%	-2.44%	-1.37%
		HSAST	-0.11%	-1.36%	-2.32%	-2.41%	-2.57%	-1.77%
	512	Lloyd	0.0044471	0.00444713	0.00444713	0.00443157	0.00443157	578.74
		Swap	48.55%	40.92%	34.75%	31.71%	28.56%	-5.77%
		HEZ	0.21%	-0.82%	-1.00%	-0.72%	-0.72%	-1.94%
		HSA	0.00%	-0.82%	-1.91%	-2.07%	-2.29%	-1.57%
		HSAST	0.32%	-1.40%	-2.11%	-2.29%	-2.51%	-1.89%
	1024	Lloyd	0.0032307	0.00322488	0.00322099	0.00320904	0.00320904	973.02
		Swap	45.64%	41.13%	36.06%	33.91%	30.72%	-5.55%
		HEZ	-0.96%	-0.99%	-1.08%	-0.72%	-0.72%	-2.32%
		HSA	-1.16%	-1.81%	-2.39%	-2.40%	-2.75%	-1.50%
		HSAST	-1.47%	-2.21%	-2.64%	-2.66%	-2.97%	-1.65%

D.7 K-Means Clustering for 20 MHz

TABLE D.6 CQI Top Mass Mode based K-Means Clustering for 20 MHz

CQI Set	N_{CT}	Method	Best Average Distortion					CPU Time
			Stage 100	Stage 250	Stage 500	Stage 750	Stage 1000	
Top 3 33596	64	Lloyd	0.0197695	0.0197354	0.0195733	0.0193381	0.0193381	16.09
		Swap	56.68%	34.22%	28.17%	28.43%	28.13%	-1.18%
		HEZ	1.24%	1.41%	-2.94%	-1.76%	-1.76%	-1.55%
		HSA	4.69%	-4.89%	-7.61%	-6.97%	-7.09%	-3.29%
		HSAST	-2.71%	-6.26%	-7.79%	-7.31%	-7.34%	-3.36%
	128	Lloyd	0.0094254	0.00941394	0.00941394	0.00941394	0.00941394	25.96
		Swap	81.38%	53.87%	38.00%	32.66%	30.44%	0.39%
		HEZ	5.89%	4.90%	4.90%	1.58%	1.58%	-1.43%
		HSA	7.35%	-0.78%	-5.04%	-7.37%	-8.06%	-0.89%
		HSAST	4.15%	-3.49%	-7.43%	-7.78%	-8.16%	-0.58%
		Lloyd	0.004542	0.00454196	0.00454196	0.00454196	0.00454196	43.44
		Swap	113.06%	69.83%	47.21%	37.86%	34.30%	1.75%

	256	HEZ	5.72%	5.72%	5.32%	5.32%	3.22%	-0.39%
		HSA	2.01%	-1.22%	-7.20%	-8.59%	-9.44%	-0.35%
		HSAST	1.70%	-1.30%	-7.07%	-7.36%	-7.87%	-0.23%
	512	Lloyd	0.0022824	0.0022772	0.00224586	0.00224586	0.00224586	75.01
		Swap	115.71%	83.31%	71.05%	59.66%	54.26%	2.51%
		HEZ	0.42%	-0.84%	-0.47%	-0.47%	-0.47%	0.83%
		HSA	2.31%	-0.69%	-2.75%	-4.12%	-5.73%	1.49%
		HSAST	0.13%	-1.58%	-3.85%	-5.42%	-5.88%	0.71%
	1024	Lloyd	0.0011522	0.00115216	0.00115216	0.00115216	0.00113828	139.07
		Swap	125.94%	114.26%	100.31%	91.05%	84.94%	-2.24%
		HEZ	-1.73%	-2.24%	-2.24%	-2.24%	-1.05%	-2.80%
		HSA	-2.10%	-3.09%	-4.39%	-5.91%	-5.19%	-3.89%
		HSAST	-2.27%	-3.12%	-4.25%	-5.42%	-4.76%	-3.98%
Top 4 144179	64	Lloyd	0.0111758	0.0110538	0.0110538	0.0110538	0.0110538	56.9
		Swap	65.51%	46.93%	38.80%	34.12%	31.19%	4.52%
		HEZ	7.10%	7.75%	4.48%	4.48%	3.38%	-3.80%
		HSA	7.43%	1.68%	-1.01%	-1.74%	-2.42%	-4.09%
		HSAST	1.06%	0.70%	-2.10%	-2.23%	-2.64%	-6.82%
	128	Lloyd	0.0067669	0.0067057	0.00658658	0.00658658	0.00645069	90.25
		Swap	78.89%	53.00%	35.74%	31.63%	32.78%	2.66%
		HEZ	4.87%	4.11%	1.28%	0.42%	2.54%	-3.39%
		HSA	1.68%	-1.86%	-0.64%	-2.15%	-0.78%	-3.58%
		HSAST	1.25%	-0.88%	-2.00%	-3.02%	-1.66%	-1.37%
	256	Lloyd	0.004161	0.00414648	0.00414648	0.00414648	0.00414648	143.03
		Swap	85.59%	55.22%	38.75%	33.61%	29.44%	1.41%
		HEZ	1.19%	0.51%	0.51%	0.51%	-0.79%	-3.45%
		HSA	1.41%	-0.28%	-2.26%	-3.51%	-4.29%	-2.98%
		HSAST	1.53%	-0.77%	-3.28%	-3.68%	-4.05%	-2.24%
	512	Lloyd	0.0027201	0.00268813	0.00268813	0.00266159	0.00266159	223.8
		Swap	65.98%	51.98%	41.30%	36.85%	32.51%	0.26%
		HEZ	-0.92%	0.10%	-1.45%	-1.17%	-1.17%	-1.30%
		HSA	-0.83%	-1.00%	-3.14%	-2.99%	-3.52%	-1.86%
		HSAST	-1.41%	-2.15%	-3.61%	-3.35%	-3.94%	-1.94%
	1024	Lloyd	0.0017489	0.00173112	0.00172216	0.00172216	0.00172216	360.7
		Swap	59.47%	55.75%	49.32%	43.63%	37.59%	-0.11%
		HEZ	-1.54%	-0.65%	-0.28%	-0.28%	-0.28%	-0.38%
		HSA	-0.91%	-1.05%	-2.14%	-3.06%	-3.51%	-1.64%
		HSAST	-1.40%	-1.37%	-1.92%	-2.83%	-3.32%	-0.73%
Top 5 206473	64	Lloyd	0.0085583	0.0085017	0.0085017	0.00839108	0.00839108	137.19
		Swap	57.46%	39.73%	32.68%	28.08%	26.42%	-1.06%
		HEZ	2.55%	-0.35%	-0.35%	0.96%	0.96%	-3.05%
		HSA	0.76%	-0.09%	-1.71%	-1.24%	-1.47%	-5.67%
		HSAST	0.20%	-1.55%	-2.84%	-1.76%	-1.76%	-6.82%
	128	Lloyd	0.0058201	0.00582013	0.00582013	0.00582013	0.00582013	208.18
		Swap	55.74%	42.60%	31.64%	26.77%	25.20%	0.96%
		HEZ	1.34%	1.23%	0.15%	0.15%	0.15%	-2.32%
		HSA	1.88%	-0.82%	-1.86%	-2.36%	-2.52%	-3.53%
		HSAST	-0.29%	-2.18%	-2.73%	-2.87%	-3.28%	-4.91%
	256	Lloyd	0.004181	0.00418103	0.00414545	0.00414545	0.00412509	320.52
		Swap	50.66%	36.74%	31.12%	27.29%	24.36%	-0.94%
		HEZ	-1.09%	-1.24%	-0.39%	-0.39%	0.10%	-1.04%
		HSA	-0.87%	-2.41%	-3.18%	-3.35%	-3.01%	-2.30%
		HSAST	-1.35%	-3.03%	-2.81%	-3.46%	-3.23%	-2.25%
	512	Lloyd	0.0030008	0.00297438	0.00297407	0.00297407	0.00296399	504.21
		Swap	48.00%	42.85%	34.77%	29.87%	27.50%	-2.73%
		HEZ	-0.79%	0.09%	-0.40%	-1.14%	-0.80%	-1.76%
		HSA	-0.48%	-1.02%	-1.83%	-2.76%	-2.84%	-1.61%
		HSAST	-1.15%	-1.80%	-2.86%	-3.51%	-3.30%	-1.91%
	1024	Lloyd	0.0021459	0.00214587	0.00214587	0.00214096	0.00214096	820.75
		Swap	49.38%	45.53%	40.26%	36.50%	33.26%	-2.24%
		HEZ	0.05%	-0.78%	-0.82%	-0.59%	-0.59%	-1.40%
		HSA	0.13%	-1.00%	-2.01%	-2.43%	-2.90%	-1.47%
		HSAST	-0.21%	-1.47%	-2.33%	-2.69%	-3.03%	-1.43%

D.8 Summary

The simulation results presented in this section are conducted for the entire set of existing bandwidths in LTE by using the following reassignment schemes in the preprocessing stage: Top3, Top4 and Top5 reassignment mass modes. For each preprocessing scheme, the performances of different sets of centers are analyzed when $N_{CT} = \{64,128,256,512,1024\}$. The Lloyd algorithm performs the local search by moving the centers to the corresponding centroids of the preprocessed CQI data points at each stage. The Swap heuristic aims to replace at each stage one center with one preprocessed CQI data point from the candidate list. This way, Swap heuristic can perform much better than Lloyd but under a longer simulation time. From these reasons, the Swap heuristic provides the highest best average distortion when compared against other heuristics since the simulation time of 1000 stages is not enough to find the global minimum solutions. Hybrid-EZ performs a swap heuristic at the beginning of each run and for the rest of the stages, the Lloyd heuristic is used. This way, Hybrid-EZ can provide better set of centers for the entire set of LTE bandwidths when compared against Lloyd and Swap heuristics. Hybrid-SA and Hybrid-SAST combines the Lloyd and Swap heuristics in such a way that only when the relative consecutive distortion is higher than a predefined minimum level, then the Lloyd stage may be performed based on the acceptance probability. Also, these approaches can avoid the local minima problems by accepting non-better solutions. In more than 70% of the considered simulations being provided in this section, the proposed hybrid-SAST outperforms the hybrid-SA heuristic from the viewpoint of the best average distortion obtained at the end of the simulation time. Being the best alternative for the k-mean clustering approaches, the sets of centers provided by the novel hybrid-SAST heuristic are used by the RBFNN hidden layers when the classification stage is performed.

Appendix E

Performance Evaluation of Clustering Algorithms for Variable Number of Centers

E.1 Appendix Outline

The performance of heuristic algorithms for k-means clustering is analysed in this section in terms of the best average distortion and the CPU execution time for different bandwidths under the variability of the number of CQI data centers. The impact of the number of centers in the average distortion performance for the proposed Hybrid-SAST heuristic algorithm is presented in Chapter 4, Sub-section 4.7.2, and this section highlights the results of other existing heuristics such as: Lloyd, single Swap, Hybrid-EZ and Hybrid-SA. All heuristic algorithms are performed over the collected sets of preprocessed CQI observations for different reassignment schemes such as: Top3, Top4 and Top5 mass modes. The entire set of simulations counts of about 72 processes and the results have been collected after more than three weeks of the distributed simulations on 18 machines. For each number of centers, system bandwidth and reassignment principle, each algorithm is launched for 1000 stages. At the end of one simulation, the best average distortion and the CPU execution time are saved and then, a new simulation starts by increasing the center set size with one CQI center.

E.2 Hybrid-SA Based K-Means Clustering

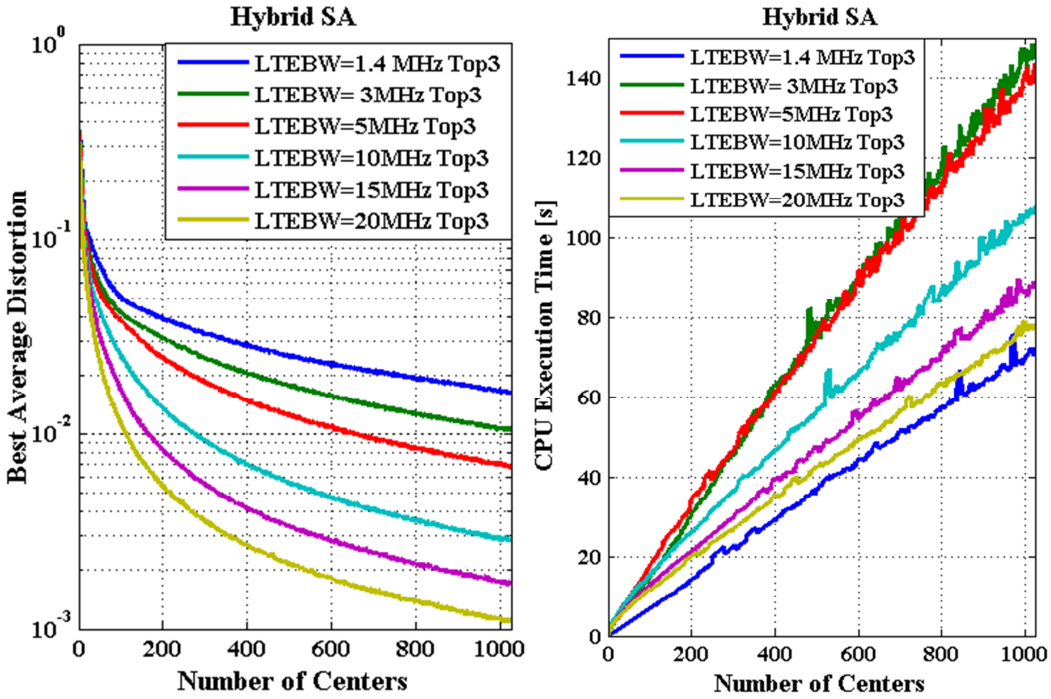


Fig. E.1 a) The Best Average Distortion and b) The CPU Execution Time for Hybrid-SA for the Top 3 CQI Mass Mode

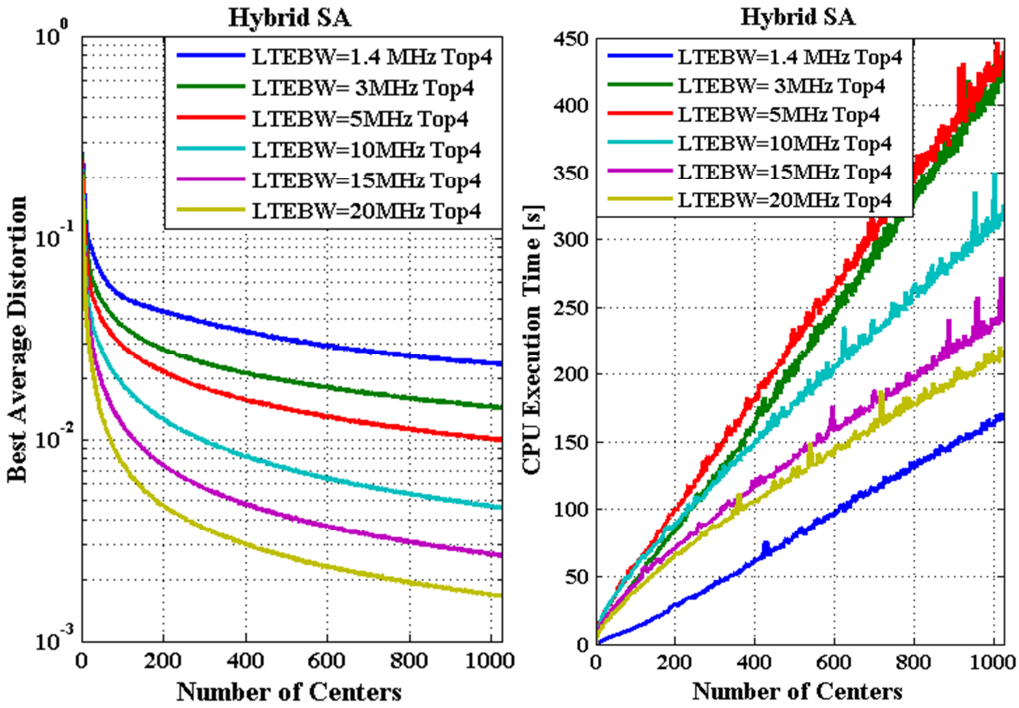


Fig. E.2 a) The Best Average Distortion and b) The CPU Execution Time for Hybrid-SA for the Top 4 CQI Mass Mode

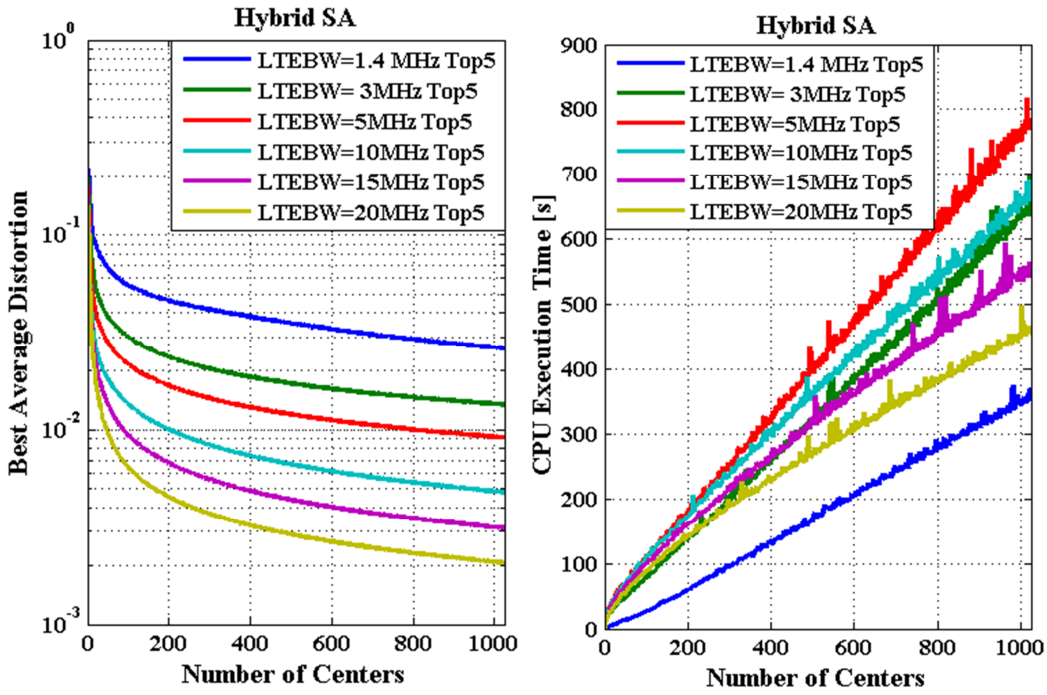


Fig. E.3 a) The Best Average Distortion and b) The CPU Execution Time for Hybrid-SA for the Top 5 CQI Mass Mode

E.3 Lloyd K-Means Clustering

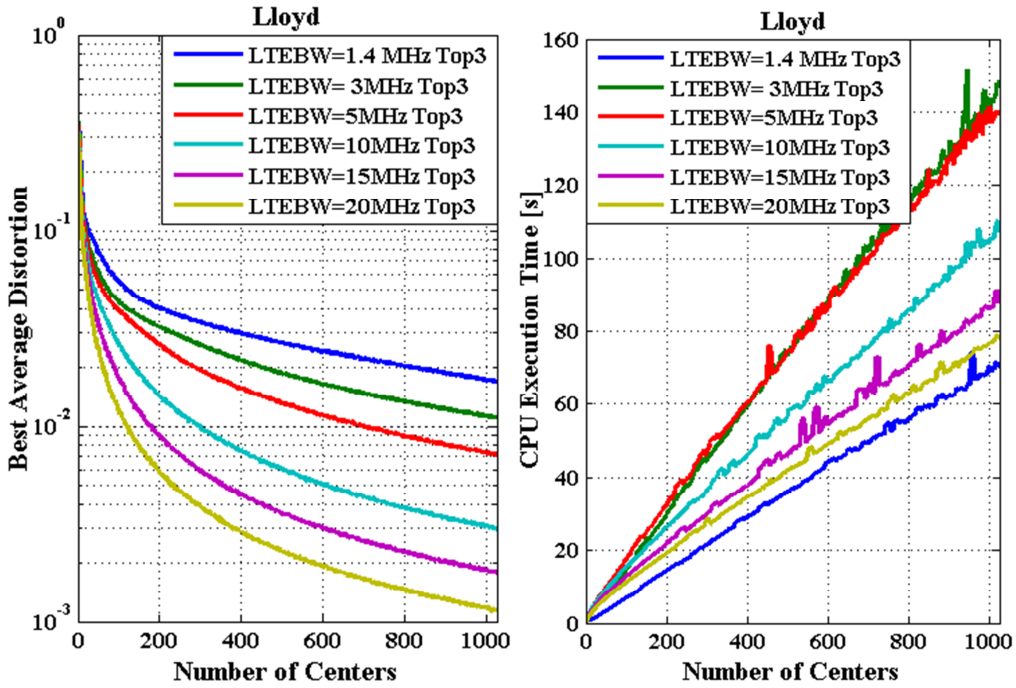


Fig. E.4 a) The Best Average Distortion and b) The CPU Execution Time for Lloyd for the Top 3 CQI Mass Mode

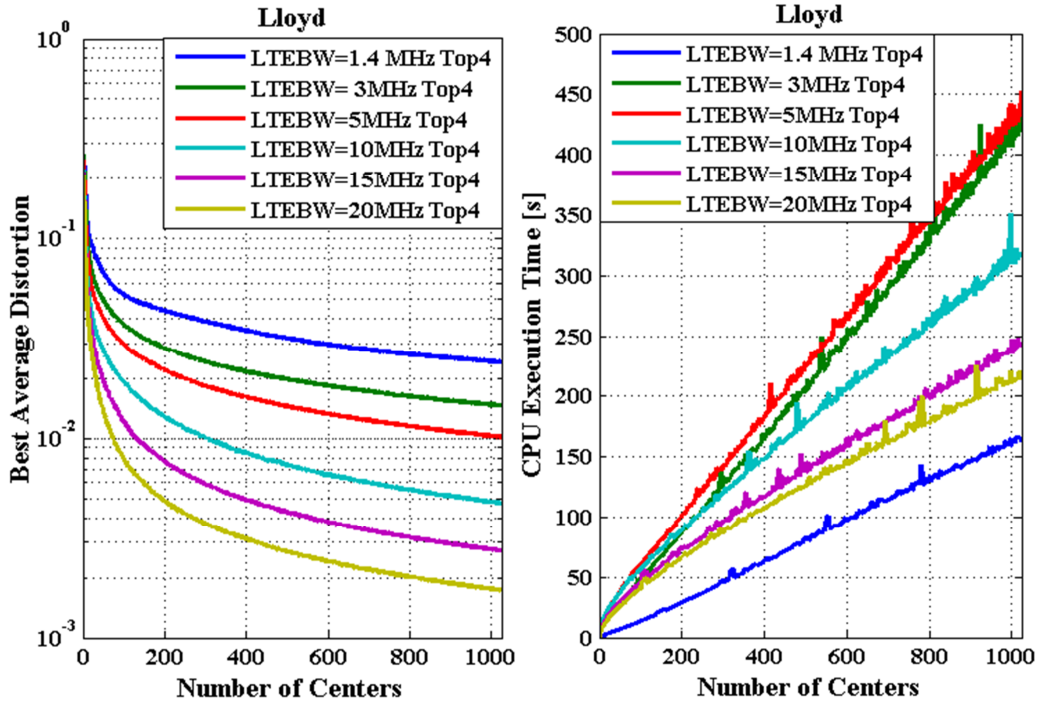


Fig. E.5 a) The Best Average Distortion and b) The CPU Execution Time for Lloyd for the Top 4 CQI Mass Mode

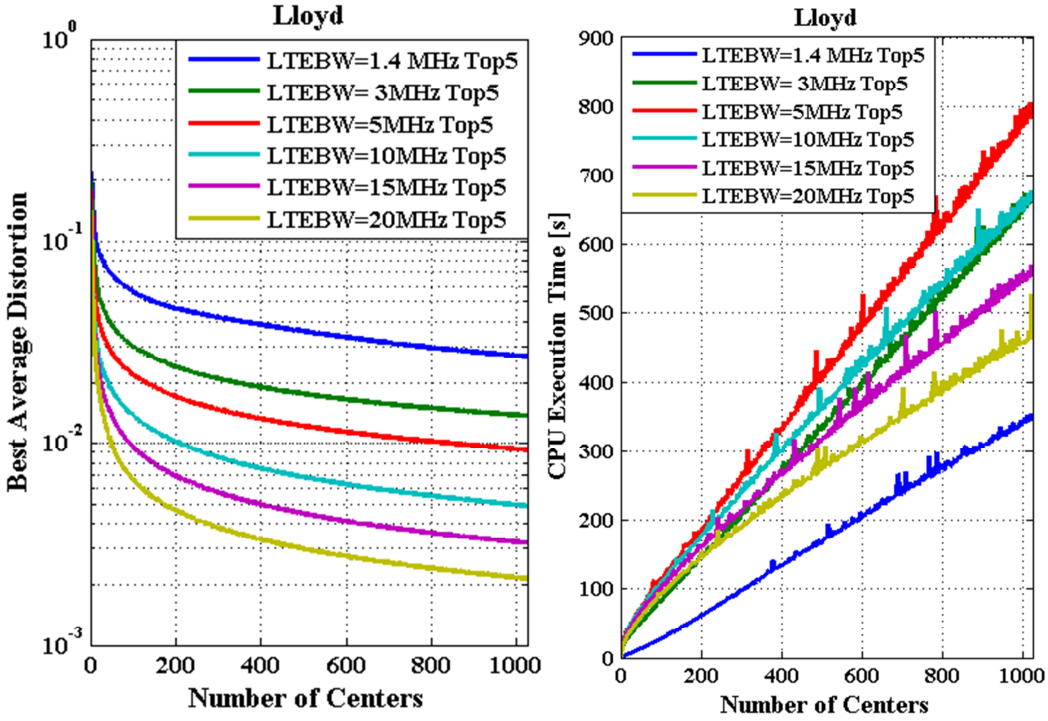


Fig. E.6 a) The Best Average Distortion and b) The CPU Execution Time for Lloyd for the Top 5 CQI Mass Mode

E.4 Swap K-Means Clustering

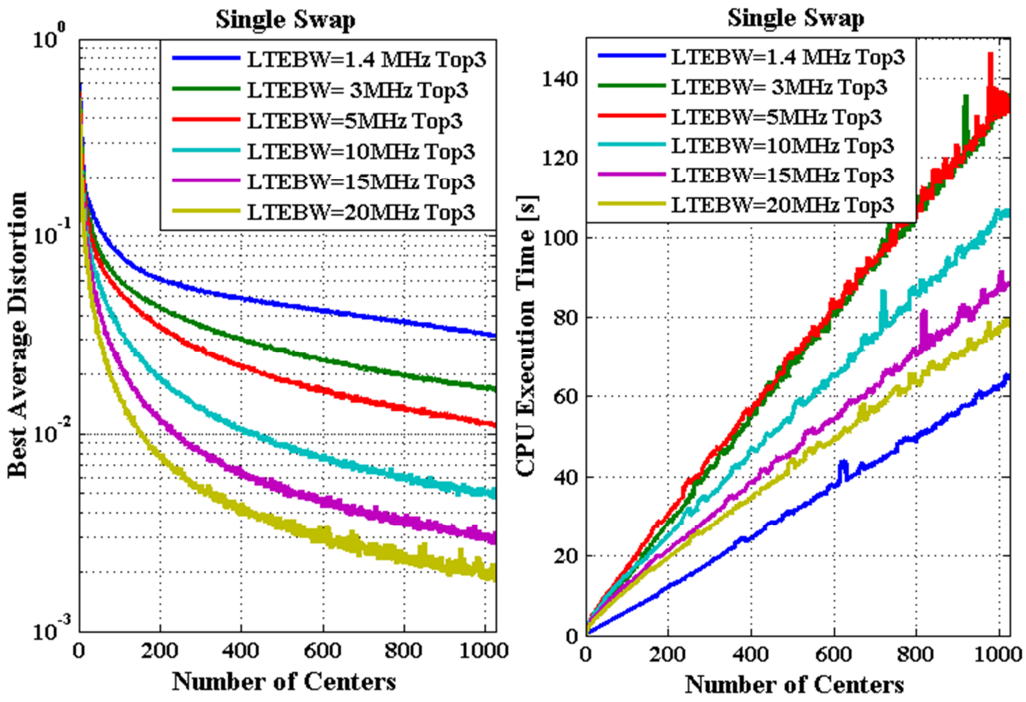


Fig. E.7 a) The Best Average Distortion and b) The CPU Execution Time for Swap for the Top 3 CQI Mass Mode

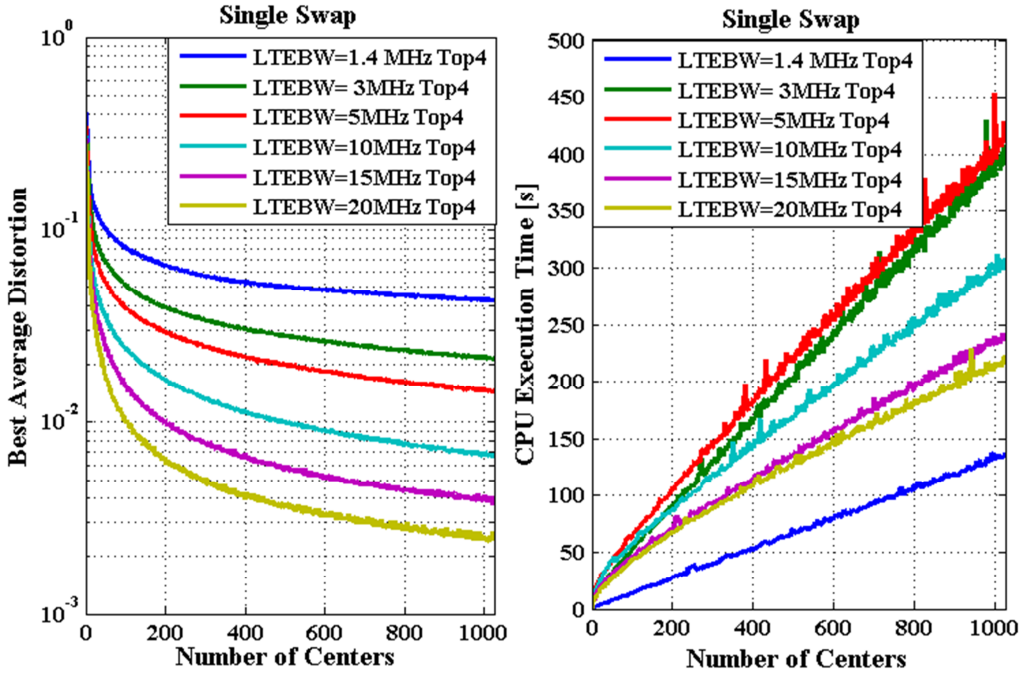


Fig. E.8 a) The Best Average Distortion and b) The CPU Execution Time for Swap for the Top 4 CQI Mass Mode

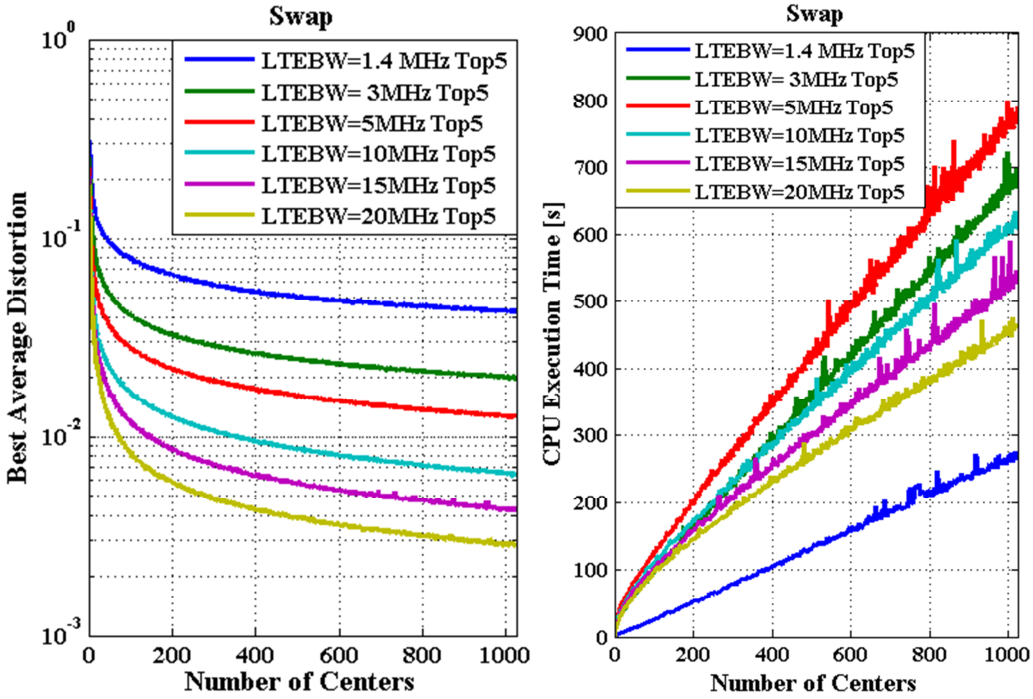


Fig. E.9 a) The Best Average Distortion and b) The CPU Execution Time for Swap for the Top 5 CQI Mass Mode

E.5 Hybrid EZ K-Means Clustering

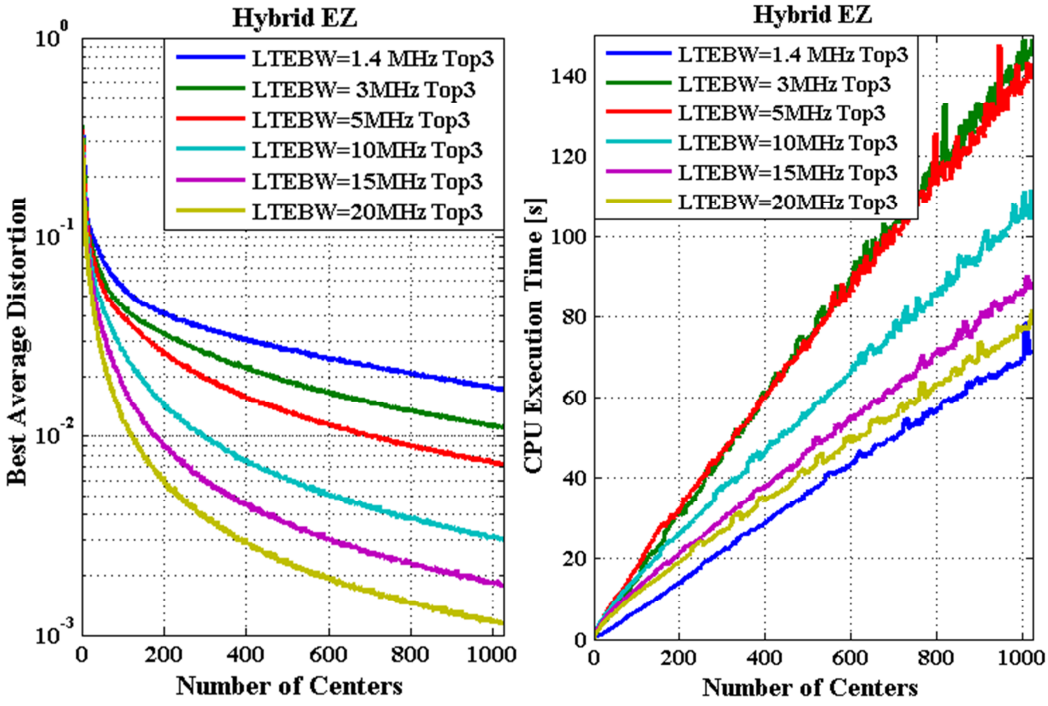


Fig. E.10 a) The Best Average Distortion and b) The CPU Execution Time for the Hybrid EZ for Top 3 CQI Mass Mode

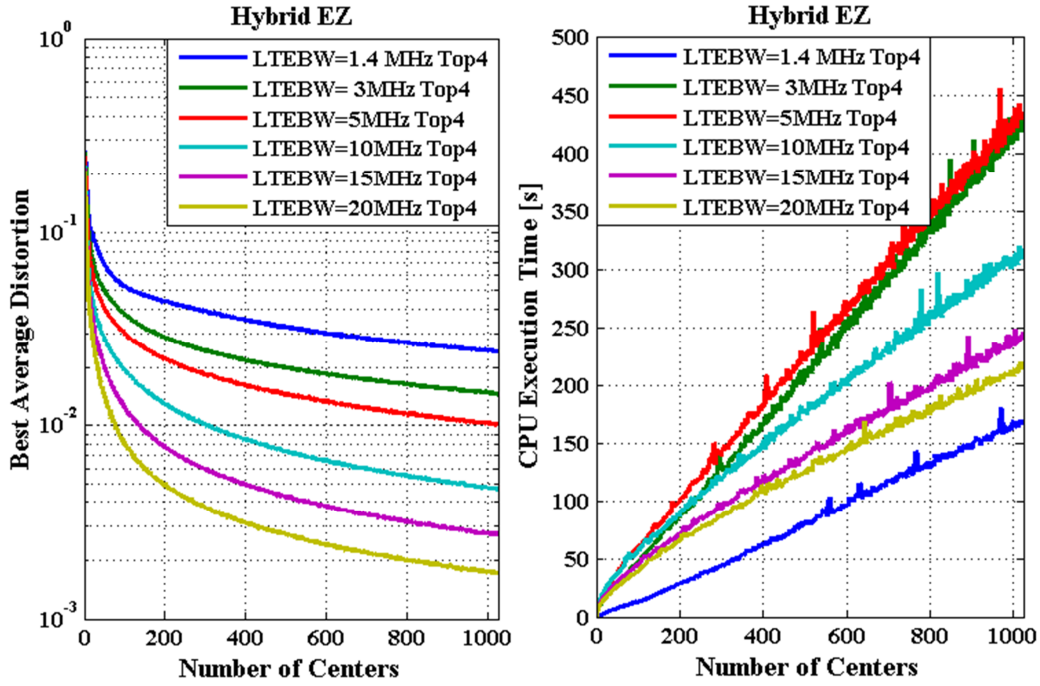


Fig. E.11 a) The Best Average Distortion and b) The CPU Execution Time for the Hybrid EZ for Top 4 CQI Mass Mode

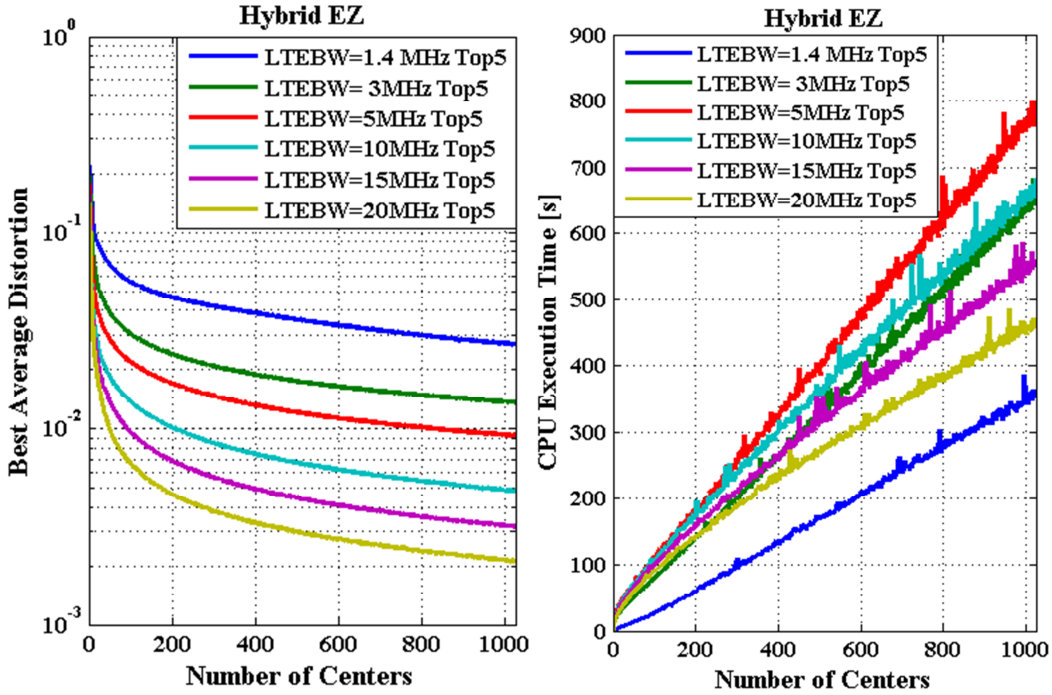


Fig. E.12 a) The Best Average Distortion and b) The CPU Execution Time for the Hybrid EZ for Top 5 CQI Mass Mode

E.6 Summary

From the viewpoints of Figures E.1.a to E.12.a, the lower the preprocessed CQI collection size is, the higher the best average distortion is for the entire set of heuristic algorithms, reassignment schemes and sets of preprocessed CQI data centers. For the system bandwidth of 1.4 MHz, the preprocessed CQI collection size registers the lowest amount of observations for all top mass reassignment principles when compared with other bandwidths. From these reasons, the average distortion is much higher than in other cases. The Lloyd heuristic can perform better for this preprocessed CQI collection if the simulation time for each set of centers is larger than 1000 stages. On the other side, the system bandwidth of 20MHz has the largest amount of preprocessed CQI observations and the best average distortion is reduced when compared with the best average distortion obtained for the collection of 1.4MHz. Basically, the higher the population size is, the higher is the probability of finding better sets of centers. When the CPU execution time is considered, the preprocessed CQI collection of 1.4 MHz followed by the preprocessed CQI collection of 20MHz indicate the best performance under the variation of the number of preprocessed CQI data centers. Due to the statistical properties of the preprocessed CQI data points for the system bandwidth of 5MHz, the CPU execution time indicates the worst performance for the entire set of algorithms and reassignment schemes.

Appendix F

Performance Evaluation of Sustainable Scheduling Policies Focusing on NGMN Fairness Requirement

F.1 Appendix Outline

This section represents an extension of Sub-section 6.2.5 from Chapter 6 and analyses the performance of the proposed scheduling policies when the AUT-EMF and AUT-MMF observations are considered in the DSR-SMOO problems focusing on NGMN fairness. The simulation results are labelled and averaged over 10 simulations in the exploitation stage. The results are represented based on the mean and STD values. At the bottom of each table, the worst and the best performances are highlighted for a more comprehensive representation. At the very basic level, the results are conducted through the mean percentage of TTIs when the scheduler stays over-fair, unfair or feasible and the mean percentage of TTIs when the exploitation rewards take different forms. If the AUT-EMF observations are considered, the scheduling policies are tested based on various CQI aggregation schemes and if the AUT-MMF observations are taken into account, the obtained policies are evaluated based on multiple windowing factors.

F.2 DSR-SMOO Focusing on the NGMN Fairness Objective with AUT-EMF Observations

The DSR-SMOO problems with the AUT-EMF observations are evaluated based on 12 configurations which can be obtained by using the reassignment top mass principles of $\{Top3, Top4, Top5\}$ and the number of centers belonging to $N_{CT} = \{64, 128, 256, 512\}$. Also, the numerical results obtained by the scheduling policies which are not considering the CQI aggregation in the state space computation are presented in Table F.1. The obtained policies are trained by using the following RL principles: Q-L, DQ-L, SARSA, QV, QV2, QVMAX, QVMAX2, ACLA, CACLA1 and CACLA2. These policies are compared with the existing techniques such as MT and AS and simple scheduling rules obtained from the generalized PF: MaxTh, MaxFair and the classical PF rule. When the performance of the proposed scheduling policies are analysed in terms of the percentage of TTIs when the controller state is $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$, then the mean percentage of TTIs and the associated deviations are presented over 10 simulations with different channel conditions. In general, the sustainable policy is the best set of scheduling rules being represented in the time domain which has the maximum percentage of feasible TTIs, the minimum percentage of unfair TTIs and the minimum STD values. In this case, the best scheduling policy is marked in green. The second best choice is marked in yellow and finally, the worst choice is denoted by the red colour. However, each table presents at the bottom the best policy from the viewpoint of the mean percentage of TTIs when the controller belongs to one of the regions $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. The simulation results for the aforementioned configurations are presented in Tables F.1 to F.13. It is very important to study the impact of the moderate and punishment rewards in the learned policies. In this sense, the mean percentage of TTIs and the STD values when the testing rewards are moderate, punishment or maximized are presented in Tables F.14 to F.26. This way, there is the possibility of monitoring how fast a sustainable scheduling policy can recover the feasible state based on the moderate and punishment rewards.

Table F.1 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ without CQI Aggregation

Mean / STD — RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.946	0.00087	0	0	0.054	0.00087
PF	1.121	0.13111	0.454	0.0764	98.426	0.17616
Max-Fair	1.114	0.13911	0.315	0.01008	98.572	0.13586
MT	10.375	0.12045	85.861	0.19283	3.764	0.13282
AS	10.742	0.13211	85.887	0.13852	3.371	0.12946
Q-L	1.917	0.18047	13.05	0.60241	85.033	0.49204
DoubleQ	2.386	0.2761	28.786	0.17126	68.828	0.14264
SARSA	2.882	0.40013	69.679	1.542	27.44	1.55255
QV	11.365	0.85216	80.337	6.69509	8.298	6.72805
QV2	60.755	0.66968	38.915	0.68463	0.33	0.02423
QVMAX	26.278	2.16499	73.171	2.18498	0.551	0.10583
QVMAX2	8.086	0.98007	22.809	1.07111	69.105	0.73658
ACLA	9.874	0.80527	82.974	1.56743	7.152	0.86744
CACLA1	11.129	0.53915	70.793	1.35498	18.077	0.9427
CACLA2	16.9	0.49729	82.734	0.58251	0.366	0.1091

Table F.2 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top3, N_{CT} = 64\}$.

Mean / STD — RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.942	0.00389	0	0	0.058	0.00389
PF	1.019	0.02499	0.491	0.1726	98.49	0.19074
Max-Fair	1.018	0.01505	0.302	0.01142	98.68	0.02324
MT	10.481	0.23213	85.816	0.2304	3.703	0.22908
AS	10.879	0.15935	85.764	0.14131	3.357	0.11217
Q-L	11.027	0.03059	86.662	0.0629	2.311	0.06811
DoubleQ	12.036	1.1916	56.955	0.51483	31.009	0.75478
SARSA	1.935	0.05811	87.96	0.85504	10.105	0.89637
QV	3.43	0.15745	90.53	1.4767	6.041	1.47477
QV2	8.703	0.48589	90.24	0.49817	1.058	0.04967
QVMAX	9.906	0.06892	88.036	0.08971	2.058	0.06383
QVMAX2	3.642	0.08023	93.797	0.12445	2.561	0.07203
ACLA	4.607	0.07356	90.878	0.15661	4.514	0.09094
CACLA1	2.739	0.05966	96.644	0.07151	0.617	0.04114
CACLA2	2.107	0.06391	96.781	0.13959	1.112	0.15841

Table F.3 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top3, N_{CT} = 128\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.893	0.09108	0	0	0.107	0.09108
PF	1.026	0.01467	0.55	0.22906	98.424	0.24095
Max-Fair	1.021	0.01086	0.306	0.01035	98.674	0.0208
MT	10.388	0.15901	85.873	0.16724	3.739	0.13063
AS	10.812	0.17623	85.763	0.09268	3.425	0.14226
Q-L	5.528	0.07359	89.598	0.21419	4.875	0.2181
DoubleQ	3.736	0.09161	80.841	0.86013	15.423	0.84364
SARSA	2.661	1.1434	87.158	1.00166	10.181	0.59622
QV	1.776	0.16175	94.567	4.01063	3.657	4.06251
QV2	3.492	0.07923	87.239	0.20279	9.269	0.15839
QVMAX	9.833	0.78522	87.66	0.66918	2.507	0.32161
QVMAX2	7.439	0.3857	91.677	0.41307	0.885	0.0646
ACLA	3.361	0.03782	92.691	0.1353	3.948	0.10482
CACLA1	3.938	0.24009	94.871	0.19319	1.191	0.27008
CACLA2	2.117	0.05756	96.157	0.34583	1.726	0.30572
Best	1.776	**QV**	96.157	**CACLA2**	0.885	**QVMAX2**
Worst	10.812	**AS**	80.841	**DoubleQ**	15.423	**DoubleQ**

Table F.4 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top3, N_{CT} = 256\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.936	0.01767	0	0.00015	0.064	0.01768
PF	1.016	0.01498	0.398	0.05782	98.586	0.06588
Max-Fair	1.014	0.0139	0.302	0.0121	98.684	0.02031
MT	10.479	0.10706	85.773	0.24076	3.749	0.15673
AS	10.773	0.1818	85.91	0.21928	3.317	0.13304
Q-L	5.913	0.13156	69.475	0.9451	24.612	0.95196
DoubleQ	10.865	0.02973	86.778	0.19394	2.356	0.1888
SARSA	1.701	0.03525	84.765	0.54675	13.534	0.56563
QV	4.014	0.19967	94.02	0.22629	1.966	0.06915
QV2	8.925	0.1623	88.591	0.19784	2.484	0.11935
QVMAX	1.702	0.03273	64.585	0.08641	33.713	0.07425
QVMAX2	2.767	0.14863	93.387	0.15454	3.846	0.08614
ACLA	4.816	0.06466	93.128	0.10175	2.057	0.04432
CACLA1	1.868	0.03529	96.543	0.25423	1.589	0.24622
CACLA2	2.064	0.03766	94.893	0.18054	3.043	0.16815
Best	1.701	**SARSA**	96.543	**CACLA1**	1.589	**CACLA1**
Worst	10.865	**DoubleQ**	64.585	**QVMAX**	33.713	**QVMAX**

Table F.5 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top3, N_{CT} = 512\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.908	0.06198	0	0.00015	0.092	0.062
PF	1.009	0.01196	0.467	0.11408	98.523	0.11232
Max-Fair	1.018	0.00949	0.305	0.01074	98.678	0.0128
MT	10.406	0.19596	85.848	0.24909	3.745	0.1211
AS	10.743	0.13799	85.85	0.15065	3.407	0.08997
Q-L	25.752	0.18719	72.215	0.18639	2.033	0.06512
DoubleQ	2.838	0.04178	44.2	0.42339	52.962	0.42891
SARSA	35.245	2.70898	58.125	2.57321	6.63	0.58291
QV	8.485	0.21003	91.127	0.21986	0.388	0.06909
QV2	3.288	0.0639	91.398	0.24416	5.314	0.20515
QVMAX	2.486	0.07372	86.174	0.61577	11.341	0.6535
QVMAX2	2.455	0.09906	93.424	0.37687	4.121	0.36414
ACLA	4.081	0.07553	94.186	0.14644	1.733	0.10627
CACLA1	5.432	0.60267	93.923	0.54724	0.645	0.12815
CACLA2	1.827	0.03298	96.484	0.07001	1.688	0.06864
Best	1.827	**CACLA2**	96.484	**CACLA2**	0.388	**QV**
Worst	35.245	**SARSA**	44.2	**DoubleQ**	52.962	**DoubleQ**

Table F.6 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top4, N_{CT} = 64\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.907	0.06873	0	0.0002	0.093	0.06876
PF	1.018	0.01738	0.459	0.08286	98.523	0.08864
Max-Fair	1.022	0.01285	0.298	0.00874	98.68	0.02065
MT	10.381	0.20216	85.812	0.21343	3.807	0.17919
AS	10.738	0.13958	85.89	0.17466	3.372	0.13866
Q-L	2.835	0.15835	76.436	0.80937	20.729	0.77524
DoubleQ	9.505	1.0554	54.824	1.18965	35.671	1.15053
SARSA	2.474	0.1262	90.843	0.51281	6.683	0.50989
QV	5.961	0.31734	91.979	0.30464	2.06	0.11587
QV2	7.949	0.51956	87.538	0.60743	4.513	0.13276
QVMAX	10.993	0.05397	87.219	0.13927	1.788	0.11166
QVMAX2	8.373	0.17181	89.44	0.17583	2.187	0.06267
ACLA	3.33	0.14351	94.455	0.13464	2.215	0.07544
CACLA1	1.771	0.04671	96.568	0.10694	1.661	0.07827
CACLA2	1.687	0.04735	95.231	0.26982	3.083	0.253
Best	1.687	**CACLA2**	96.568	**CACLA1**	1.661	**CACLA1**
Worst	10.993	**QVMAX**	54.824	**DoubleQ**	35.671	**DoubleQ**

Table F.7 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top4, N_{CT} = 128\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.938	0.01032	0	0.00023	0.062	0.01028
PF	1.014	0.01893	0.445	0.0706	98.542	0.08371
Max-Fair	1.018	0.01632	0.305	0.00937	98.677	0.02112
MT	10.418	0.13004	85.739	0.12623	3.843	0.12514
AS	10.703	0.08568	85.928	0.19623	3.369	0.14822
Q-L	8.203	0.05357	65.693	0.1553	26.105	0.12291
DoubleQ	5.695	0.18143	67.828	0.40111	26.477	0.48812
SARSA	2.023	0.08042	90.737	0.42022	7.24	0.40386
QV	6.617	0.13626	88.738	1.05419	4.645	1.05235
QV2	4.021	0.08537	91.013	0.61475	4.967	0.61583
QVMAX	2.606	0.06601	86.874	0.22169	10.519	0.16636
QVMAX2	4.381	0.18825	93.124	0.19898	2.495	0.05439
ACLA	4.761	0.11369	92.739	0.13669	2.5	0.05239
CACLA1	1.733	0.03935	94.537	0.19216	3.73	0.19112
CACLA2	2.835	0.07638	96.522	0.08665	0.643	0.04174
Best	1.733	**CACLA1**	96.522	**CACLA2**	0.643	**CACLA2**
Worst	10.703	**AS**	65.693	**Q-L**	26.477	**DoubleQ**

Table F.8 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top4, N_{CT} = 256\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.928	0.03727	0	0.0003	0.072	0.03698
PF	1.004	0.02859	0.436	0.15793	98.56	0.18172
Max-Fair	1.012	0.01728	0.301	0.01284	98.687	0.02585
MT	10.337	0.20164	85.939	0.33183	3.724	0.16927
AS	10.824	0.17975	85.805	0.17263	3.37	0.10554
Q-L	1.64	0.05309	40.017	0.28652	58.343	0.30365
DoubleQ	6.775	0.11254	88.931	0.16595	4.294	0.08693
SARSA	1.721	0.06151	87.372	0.64825	10.907	0.66603
QV	2.189	0.04291	92.51	0.15792	5.302	0.14868
QV2	11.035	0.09786	85.071	0.3321	3.894	0.28669
QVMAX	2.164	0.20438	90.423	0.35447	7.413	0.24023
QVMAX2	2.713	0.28809	92.255	0.33559	5.032	0.23276
ACLA	3.896	0.0869	92.885	0.15305	3.22	0.09219
CACLA1	1.877	0.20406	97.337	0.24525	0.785	0.13381
CACLA2	1.885	0.07781	97.536	0.06553	0.579	0.05072
Best	1.64	**Q-L**	97.536	**CACLA2**	0.579	**CACLA2**
Worst	11.035	**QV2**	40.017	**Q-L**	58.343	**Q-L**

Table F.9 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top4, N_{CT} = 512\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.918	0.04439	0	0	0.082	0.04439
PF	1.011	0.01033	0.4	0.0733	98.589	0.07452
Max-Fair	1.016	0.0073	0.296	0.0106	98.688	0.01619
MT	10.437	0.12976	85.766	0.20242	3.797	0.13779
AS	10.803	0.1927	85.866	0.20628	3.331	0.10987
Q-L	12.689	0.7197	77.903	0.63041	9.408	0.36085
DoubleQ	7.438	0.05448	74.674	0.21358	17.888	0.21439
SARSA	1.683	0.03558	83.741	0.86286	14.577	0.87975
QV	5.253	0.1344	89.724	0.23955	5.023	0.16935
QV2	3.122	0.07429	93.209	0.09216	3.669	0.06816
QVMAX	5.803	0.09109	89.969	0.1577	4.228	0.10745
QVMAX2	2.082	0.06859	95.506	0.09126	2.412	0.05792
ACLA	4.183	0.09122	92.06	0.10818	3.758	0.05564
CACLA1	1.801	0.03746	93.431	0.33599	4.768	0.34304
CACLA2	1.879	0.04115	97.755	0.04883	0.365	0.02712
Best	1.801	**CACLA1**	97.755	**CACLA2**	0.365	**CACLA2**
Worst	12.689	**Q-L**	77.903	**Q-L**	17.888	**DoubleQ**

Table F.10 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top5, N_{CT} = 64\}$.

Mean / STD / RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.893	0.08815	0	0.00015	0.107	0.08818
PF	1.025	0.01061	0.486	0.1306	98.489	0.12874
Max-Fair	1.023	0.01021	0.309	0.00854	98.668	0.01327
MT	10.369	0.15212	85.778	0.21744	3.853	0.12035
AS	10.74	0.12068	85.858	0.11462	3.402	0.14183
Q-L	6.253	0.07815	79.86	0.92724	13.887	0.91154
DoubleQ	9.821	0.07011	86.936	0.14573	3.243	0.09017
SARSA	2.535	0.08985	92.779	0.25657	4.686	0.30914
QV	6.069	0.35501	92.486	0.36851	1.445	0.14033
QV2	3.16	0.27971	93.721	0.40135	3.118	0.34358
QVMAX	4.405	0.06985	90.987	0.16804	4.609	0.11316
QVMAX2	1.791	0.04816	94.218	0.13984	3.991	0.15635
ACLA	4.984	0.25287	93.581	0.25207	1.434	0.13695
CACLA1	1.768	0.02576	93.646	0.48168	4.586	0.482
CACLA2	1.692	0.02673	94.643	0.36988	3.666	0.37766
Best	1.692	**CACLA2**	94.643	**CACLA2**	1.434	**ACLA**
Worst	10.74	**AS**	79.86	**Q-L**	13.887	**Q-L**

Table F.11 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top5, N_{CT} = 128\}$.

Mean / STD — RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.934	0.01938	0	0.0002	0.066	0.01943
PF	1.022	0.02062	0.458	0.11888	98.52	0.1326
Max-Fair	1.025	0.00956	0.303	0.01384	98.672	0.01755
MT	10.507	0.24967	85.765	0.23624	3.728	0.13009
AS	10.877	0.13236	85.752	0.19622	3.37	0.10447
Q-L	11.36	0.11854	85.803	0.12163	2.837	0.07117
DoubleQ	7.337	0.50084	88.737	0.52874	3.926	0.16175
SARSA	5.849	0.80708	90.479	0.77421	3.673	0.11538
QV	1.947	0.0572	93.253	0.17911	4.8	0.17013
QV2	5.533	0.44704	88.16	0.47962	6.308	0.26572
QVMAX	3.65	0.05075	88.172	0.56324	8.178	0.54776
QVMAX2	3.623	0.08057	92.994	0.11988	3.383	0.08677
ACLA	3.364	0.06943	93.32	0.19736	3.316	0.14613
CACLA1	1.877	0.04486	94.131	0.24679	3.992	0.26313
CACLA2	1.735	0.04733	96.452	0.30938	1.813	0.33937
Best	1.735	**CACLA2**	96.452	**CACLA2**	1.813	**CACLA2**
Worst	11.36	**Q-L**	85.752	**AS**	8.178	**QVMAX**

Table F.12 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top5, N_{CT} = 256\}$.

Mean / STD — RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.92	0.03141	0.006	0.01842	0.073	0.02919
PF	1.011	0.01609	0.473	0.18242	98.515	0.18851
Max-Fair	1.014	0.0104	0.303	0.01003	98.683	0.01335
MT	10.356	0.13359	85.852	0.22388	3.792	0.15694
AS	10.755	0.11743	85.84	0.1718	3.405	0.08038
Q-L	10.423	0.43939	84.404	0.47774	5.172	0.13462
DoubleQ	1.67	0.06002	44.97	0.93554	53.36	0.95545
SARSA	2.744	0.14836	87.698	0.2399	9.558	0.18629
QV	5.807	0.14536	92.051	0.20481	2.142	0.09307
QV2	5.039	0.10703	93.033	0.12407	1.929	0.11044
QVMAX	6.578	0.39447	89.562	0.36654	3.86	0.08501
QVMAX2	6.328	0.85447	88.302	0.92466	5.369	0.12488
ACLA	3.094	0.10916	93.787	0.76308	3.119	0.82672
CACLA1	1.935	0.05241	96.552	0.08909	1.514	0.10303
CACLA2	1.978	0.30193	96.304	0.36273	1.718	0.09624
Best	1.67	**DoubleQ**	96.552	**CACLA1**	1.514	**CACLA1**
Worst	10.755	**AS**	44.97	**DoubleQ**	53.36	**DoubleQ**

Table F.14 Percentages of TTIs (Mean and STD)[%] for the Scheduler States when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$. CQI Aggregation Scheme: $\{Top5, N_{CT} = 512\}$.

Mean / STD — RL Alg.	Mean $\mathcal{S}_t^C \in \mathcal{UFF}$	STD $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean $\mathcal{S}_t^C \in \mathcal{FAF}$	STD $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean $\mathcal{S}_t^C \in \mathcal{OFF}$	STD $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	99.942	0.00629	0	0.00015	0.058	0.00631
PF	1.02	0.0121	0.477	0.06741	98.503	0.07415
Max-Fair	1.019	0.0109	0.304	0.00541	98.677	0.01036
MT	10.401	0.14871	85.874	0.21735	3.725	0.12425
AS	10.816	0.19519	85.854	0.1674	3.33	0.06676
Q-L	2.791	0.05343	86.249	0.0917	10.96	0.05994
DoubleQ	4.227	0.06393	90.151	0.29866	5.622	0.26243
SARSA	8.073	0.71867	89.828	0.69827	2.099	0.06024
QV	1.716	0.03175	95.647	0.28936	2.637	0.31038
QV2	2.11	0.06149	95.089	0.08498	2.801	0.0564
QVMAX	2.001	0.13136	88.491	0.34866	9.508	0.36047
QVMAX2	3.982	0.12078	94.547	0.13568	1.471	0.04077
ACLA	7.301	0.11496	91.481	0.13014	1.218	0.03963
CACLA1	1.894	0.20895	95.656	0.22229	2.45	0.12531
CACLA2	1.754	0.03239	97.416	0.09691	0.83	0.10812
Best	1.716	**QV**	97.416	**CACLA2**	0.83	**CACLA2**
Worst	10.816	**AS**	85.854	**AS**	10.96	**Q-L**

Table F.13 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. No CQI Aggregation Scheme

Reward Type / RL Alg.	Punish Reward Mean [%] $\mathcal{RW}_t^F = -1$	Punish Reward STD [%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean [%] $\mathcal{RW}_t^F = 1$	Max. Reward STD [%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean [%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD [%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	8.671	0.0721	13.05	0.60237	78.279	0.56518
DoubleQ	6.64	0.15272	28.783	0.17089	64.577	0.09261
SARSA	15.226	0.38348	69.679	1.542	15.095	1.56436
QV	10.167	6.68875	80.313	6.69336	9.52	0.62166
QV2	11.602	1.64929	38.895	0.68237	49.502	1.52263
QVMAX	11.13	0.59894	73.107	2.18402	15.763	1.62629
QVMAX2	8.556	0.60245	22.809	1.07114	68.636	0.88759
ACLA	7.631	0.49665	82.958	1.56752	9.411	1.099
CACLA1	7.156	0.44025	70.78	1.35587	22.064	1.09701
CACLA2	12.592	0.3243	82.672	0.57665	4.736	0.26101
Best	6.64	**DoubleQ**	82.958	**ACLA**	4.736	**CACLA2**
Worst	15.226	**SARSA**	13.05	**Q-L**	78.279	**Q-L**

Table F.15 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top3, N_{CT} = 64\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	7.151	0.10075	86.599	0.06205	6.25	0.14959
DoubleQ	8.19	0.21456	56.95	0.51507	34.86	0.6845
SARSA	1.162	0.05018	87.959	0.8543	10.88	0.85894
QV	1.829	0.23002	90.51	1.47623	7.661	1.32109
QV2	1.773	0.07201	90.188	0.49897	8.039	0.43726
QVMAX	6.703	0.09154	87.972	0.09089	5.324	0.12452
QVMAX2	1.415	0.04599	93.774	0.12639	4.811	0.08508
ACLA	4.939	0.09111	90.852	0.15676	4.209	0.0947
CACLA1	0.784	0.03629	96.624	0.07212	2.593	0.05166
CACLA2	1.353	0.10329	96.779	0.14004	1.868	0.05107
Best	1.162	**SARSA**	96.779	**CACLA2**	1.868	**CACLA2**
Worst	8.19	**DoubleQ**	56.95	**DoubleQ**	34.86	**DoubleQ**

Table F.16 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top3, N_{CT} = 128\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	2.767	0.04641	89.56	0.21336	7.673	0.19474
DoubleQ	1.251	0.03939	80.804	0.85899	17.946	0.86492
SARSA	2.279	0.1968	87.157	1.00148	10.563	0.83595
QV	1.143	0.40671	94.566	4.01068	4.291	3.62406
QV2	3.104	0.06461	87.231	0.2024	9.664	0.14794
QVMAX	4.571	0.30295	87.625	0.67039	7.804	0.42425
QVMAX2	3.524	0.17472	91.642	0.41213	4.834	0.32073
ACLA	3.118	0.08998	92.662	0.13681	4.22	0.06164
CACLA1	2.276	0.10221	94.851	0.19242	2.873	0.09963
CACLA2	1.295	0.14763	96.154	0.34712	2.552	0.22866
Best	1.143	**QV**	96.154	**CACLA2**	2.552	**CACLA2**
Worst	4.571	**QVMAX**	80.804	**DoubleQ**	17.946	**DoubleQ**

Table F.17 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top3, N_{CT} = 256\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	9.969	0.31881	69.432	0.94377	20.599	0.6404
DoubleQ	7.252	0.06608	86.714	0.19475	6.034	0.20067
SARSA	8.603	0.36256	84.765	0.54683	6.633	0.24062
QV	0.861	0.04474	93.98	0.22811	5.159	0.20674
QV2	5.88	0.11571	88.541	0.19873	5.579	0.16385
QVMAX	13.197	0.14592	64.585	0.0865	22.218	0.15882
QVMAX2	1.075	0.09272	93.379	0.15482	5.546	0.08336
ACLA	1.716	0.03772	93.099	0.10269	5.185	0.08894
CACLA1	0.733	0.10977	96.542	0.25431	2.725	0.15447
CACLA2	1.956	0.1414	94.893	0.18061	3.152	0.07486
Best	0.733	**CACLA1**	96.542	**CACLA1**	2.725	**CACLA1**
Worst	13.197	**QVMAX**	64.585	**QVMAX**	22.218	**QVMAX**

Table F.18 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top3, N_{CT} = 512\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	10.339	0.14222	72.145	0.18806	17.516	0.13967
DoubleQ	1.701	0.08042	44.186	0.42474	54.112	0.38507
SARSA	4.366	0.56457	58.119	2.57263	37.515	2.73191
QV	4.71	0.14053	91.047	0.22213	4.243	0.16061
QV2	1.964	0.06582	91.38	0.24557	6.656	0.19221
QVMAX	4.913	0.26864	86.153	0.61471	8.934	0.36211
QVMAX2	1.136	0.15069	93.421	0.37676	5.443	0.24279
ACLA	2.357	0.08409	94.156	0.14745	3.487	0.06955
CACLA1	2.573	0.24771	93.908	0.54855	3.519	0.32842
CACLA2	1.108	0.02911	96.484	0.07015	2.408	0.07479
Best	1.108	**CACLA2**	96.484	**CACLA2**	2.408	**CACLA2**
Worst	10.339	**Q-L**	44.186	**DoubleQ**	54.112	**DoubleQ**

Table F.19 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top4, N_{CT} = 64\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	6.636	0.07334	76.431	0.80902	16.933	0.77513
DoubleQ	7.167	0.2381	54.816	1.18984	38.017	1.14628
SARSA	1.581	0.13339	90.826	0.51145	7.594	0.48557
QV	4.153	0.33471	91.931	0.30598	3.917	0.11773
QV2	2.521	0.10477	87.505	0.60749	9.974	0.59368
QVMAX	7.24	0.09625	87.154	0.1393	5.606	0.10325
QVMAX2	5.777	0.169	89.378	0.17883	4.845	0.07709
ACLA	1.953	0.09226	94.419	0.13695	3.628	0.07935
CACLA1	0.742	0.03755	96.567	0.10724	2.691	0.08177
CACLA2	1.93	0.12055	95.23	0.27008	2.84	0.16177
Best	1.581	**SARSA**	96.567	**CACLA1**	2.84	**CACLA1**
Worst	7.167	**DoubleQ**	54.816	**DoubleQ**	38.017	**DoubleQ**

Table F.20 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top4, N_{CT} = 128\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	11.688	0.0819	65.632	0.15788	22.679	0.13702
DoubleQ	5.723	0.18558	67.818	0.40011	26.459	0.24987
SARSA	4.254	0.10576	90.733	0.42084	5.013	0.38816
QV	3.961	0.31047	88.704	1.05309	7.335	0.79167
QV2	2.808	0.07026	90.988	0.61411	6.203	0.61824
QVMAX	6.929	0.06087	86.868	0.22179	6.203	0.26404
QVMAX2	2.738	0.12306	93.076	0.19834	4.186	0.10542
ACLA	2.87	0.0861	92.712	0.13833	4.418	0.08189
CACLA1	0.706	0.04052	94.536	0.19203	4.758	0.1589
CACLA2	1.455	0.06726	96.521	0.08702	2.024	0.05124
Best	0.706	**CACLA1**	96.521	**CACLA2**	2.024	**CACLA2**
Worst	11.688	**Q-L**	65.632	**Q-L**	26.459	**DoubleQ**

Table F.21 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top4, N_{CT} = 256\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	13.772	0.13536	40.017	0.2865	46.211	0.26377
DoubleQ	3.296	0.05351	88.893	0.16571	7.811	0.13206
SARSA	5.775	0.32034	87.371	0.64831	6.854	0.38763
QV	1.796	0.09971	92.507	0.15777	5.697	0.09394
QV2	7.448	0.07519	85.008	0.33094	7.543	0.29119
QVMAX	2.033	0.09604	90.422	0.35432	7.545	0.3499
QVMAX2	1.581	0.15483	92.251	0.33558	6.168	0.19881
ACLA	3.838	0.07917	92.864	0.15263	3.298	0.10508
CACLA1	0.754	0.17296	97.336	0.24501	1.909	0.10527
CACLA2	1.162	0.05722	97.517	0.06658	1.321	0.01958
Best	0.754	**CACLA1**	97.517	**CACLA2**	1.321	**CACLA2**
Worst	13.772	**Q-L**	40.017	**Q-L**	46.211	**Q-L**

Table F.22 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top4, N_{CT} = 512\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	6.037	0.08961	77.846	0.63228	16.118	0.6991
DoubleQ	8.677	0.06348	74.652	0.21353	16.67	0.19547
SARSA	10.126	0.49581	83.74	0.86269	6.134	0.46605
QV	3.966	0.096	89.718	0.24161	6.316	0.17037
QV2	1.617	0.03742	93.194	0.09275	5.189	0.08347
QVMAX	3.112	0.04102	89.937	0.15834	6.951	0.14144
QVMAX2	0.717	0.07866	95.5	0.09285	3.784	0.04355
ACLA	4.085	0.07237	92.03	0.10898	3.885	0.05671
CACLA1	1.337	0.14356	93.43	0.33606	5.234	0.20267
CACLA2	1.154	0.0406	97.754	0.04907	1.092	0.01638
Best	0.717	**QVMAX2**	97.754	**CACLA2**	1.092	**CACLA2**
Worst	10.126	**SARSA**	74.652	**DoubleQ**	16.67	**DoubleQ**

Table F.23 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top5, N_{CT} = 64\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	4.802	0.07278	79.833	0.92762	15.365	0.89234
DoubleQ	6.082	0.05295	86.89	0.14698	7.029	0.14643
SARSA	3.909	0.12654	92.758	0.25542	3.332	0.15009
QV	3.872	0.24014	92.429	0.36909	3.699	0.16532
QV2	1.676	0.17231	93.689	0.40392	4.635	0.34092
QVMAX	3.641	0.0837	90.963	0.16921	5.395	0.10954
QVMAX2	0.423	0.04971	94.217	0.1396	5.36	0.10945
ACLA	2.594	0.12273	93.543	0.25459	3.863	0.14673
CACLA1	1.681	0.20241	93.645	0.48171	4.674	0.29111
CACLA2	1.864	0.15314	94.642	0.36978	3.493	0.22048
Best	0.423	**QVMAX2**	94.642	**CACLA2**	3.332	**SARSA**
Worst	6.082	**DoubleQ**	79.833	**Q-L**	15.365	**Q-L**

Table F.24 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top5, N_{CT} = 128\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	7.397	0.06936	85.737	0.12203	6.867	0.10303
DoubleQ	4.262	0.11696	88.689	0.52619	7.05	0.53197
SARSA	4.413	0.08846	90.453	0.7734	5.134	0.8155
QV	3.134	0.06452	93.25	0.179	3.616	0.13664
QV2	3.897	0.34	88.11	0.48293	7.993	0.2767
QVMAX	1.651	0.03439	88.145	0.5624	10.204	0.55223
QVMAX2	1.706	0.06377	92.964	0.12152	5.33	0.07859
ACLA	2.77	0.08692	93.288	0.19613	3.942	0.12861
CACLA1	1.116	0.07234	94.13	0.24724	4.755	0.19146
CACLA2	1.968	0.13273	96.451	0.30931	1.581	0.18172
Best	1.116	**CACLA1**	96.451	**CACLA2**	1.581	**CACLA2**
Worst	7.397	**Q-L**	85.737	**Q-L**	10.204	**QVMAX**

Table F.25 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top5, N_{CT} = 256\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	2.527	0.04833	84.359	0.4811	13.114	0.48699
DoubleQ	11.87	0.39779	44.97	0.93551	43.16	0.69354
SARSA	6.723	0.07814	87.691	0.23936	5.585	0.17866
QV	2.805	0.04204	92.019	0.20507	5.176	0.17677
QV2	2.454	0.05576	92.996	0.12475	4.55	0.10119
QVMAX	4.311	0.06584	89.533	0.36787	6.156	0.3857
QVMAX2	1.841	0.06183	88.289	0.92497	9.87	0.92123
ACLA	2.147	0.24637	93.759	0.76351	4.094	0.53552
CACLA1	0.671	0.05707	96.55	0.08916	2.779	0.05017
CACLA2	1.553	0.31326	96.304	0.36256	2.144	0.08397
Best	0.671	**CACLA1**	96.55	**CACLA1**	2.144	**CACLA2**
Worst	11.87	**DoubleQ**	44.97	**DoubleQ**	43.16	**DoubleQ**

Table F.26 Percentages of TTIs (Mean and STD) [%] in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. CQI Aggregation: $\{Top5, N_{CT} = 512\}$.

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
Q-L	7.029	0.03648	86.242	0.09151	6.729	0.0728
DoubleQ	2.061	0.04521	90.119	0.29949	7.82	0.26028
SARSA	4.348	0.07385	89.792	0.69671	5.86	0.65372
QV	0.778	0.1154	95.647	0.28939	3.574	0.17886
QV2	0.702	0.04274	95.087	0.08501	4.211	0.05348
QVMAX	1.016	0.07616	88.489	0.3485	10.494	0.31033
QVMAX2	1.675	0.05692	94.516	0.13384	3.809	0.1011
ACLA	4.304	0.05933	91.419	0.13235	4.277	0.0971
CACLA1	1.015	0.19324	95.655	0.22213	3.33	0.08056
CACLA2	1.817	0.04323	97.416	0.09686	0.768	0.05996
Best	0.702	**QV2**	97.416	**CACLA2**	0.768	**CACLA2**
Worst	7.029	**Q-L**	86.242	**Q-L**	10.494	**QVMAX**

F.3 DSR-SMOO Focusing on the NGMN Fairness Objective with AUT-MMF Observations

This sub-section provides the extensive results of Sub-section 6.2.5.3 from Chapter 6 and presents the performance of the sustainable scheduling policies when the AUT-MMF observations are used in the DSR-SMOO instantaneous problems. One CQI aggregation scheme is used for the entire set of simulations such as $(Top3, N_{CT} = 64)$. The mean percentages of TTIs and the STD values are presented when the controller stays in one of the state $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ at each TTI t for the following set of static windowing factors $\rho \in [2; 5.5]$ with a factor step of 0.25. For this set of simulations, the scheduling policies are obtained by using the following RL approaches: QV2, QVMAX2, ACLA, CACLA1 and CACLA2. The obtained policies are compared against the existing AS and MT techniques with the AUT-MMF observations and against the scheduling rules such as MaxTh, MaxFair and PF. The best policies are highlighted at the bottom of each table when the scheduler controller state is $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$.

Table F.27 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 2.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	70.135	1.57192	0	0	29.865	1.57192
PF	51.316	2.79237	34.806	2.6243	13.878	0.29581
Max-Fair	16.901	0.92515	25.582	2.32599	57.517	1.53922
MT	49.225	2.72378	36.366	2.2273	14.409	0.94952
AS	22.337	1.36117	55.861	1.45501	21.802	0.82854
QV2	21.094	1.33106	58.077	1.80445	20.829	0.70554
QVMAX2	16.448	1.22227	60.087	2.02369	23.465	0.95232
ACLA	16.135	1.10536	56.676	2.09518	27.189	1.25932
CACLA1	17.38	1.22973	59.137	1.88737	23.483	0.90952
CACLA2	20.229	1.41191	59.904	1.58696	19.867	0.40364
Best	17.38	**CACLA1**	60.087	**QVMAX2**	19.867	**CACLA2**
Worst	49.225	**MT**	36.366	**MT**	27.189	**ACLA**

Tables F.27 to F.41 indicate the mean percentages of TTIs when the scheduler is unfair, over-fair and feasible for the considered set of scheduling policies and windowing factors. The mean percentages of TTIs when the rewards take different values are highlighted in Tables F.42 to F.56 for the same range of windowing factors. The optimum range of windowing factors is reached when the mean percentage of feasible TTIs is maximized and when the number of punishment and moderate rewards is minimized. At the same time, the STD values should respect some upper bounds in order to sustain the learned scheduling policies. In fact, the STD values show if the learned scheduling policies are the subject of the over-fitting or under-fitting problems. When the STD is very high, the learned policy is not suitable to be applied due to the fact that the trained MLPNN weights suffer from the over-fitting or under-fitting problems and these structures cannot take optimal decisions when the LTE environment changes substantially. For the general case, the STD should be less than one in order to assure the optimality in selecting proper scheduling rules for the obtained set of sustainable scheduling policies. If the scheduling policy indicates high percentage of feasible TTIs and very high STD values, it is preferable to exploit those policies with lower STD values.

Table F.28 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 2.25)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	73.782	1.77989	0	0	26.218	1.77989
PF	49	3.02509	41.095	2.97692	9.905	0.34267
Max-Fair	13.84	0.82252	52.779	2.01774	33.381	1.37458
MT	43.481	2.81758	45.054	2.67639	11.465	0.45988
AS	17.485	1.99231	67.588	1.93565	14.927	0.26653
QV2	13.024	1.11207	70.909	1.23149	16.068	0.41063
QVMAX2	12.831	1.02417	71.477	1.18501	15.692	0.42746
ACLA	18.094	1.24573	68.067	1.41332	13.839	0.34658
CACLA1	13.318	1.10552	70.5	1.1968	16.183	0.37552
CACLA2	16.495	1.54921	70.189	1.53622	13.316	0.27442
Best	12.831	**QVMAX2**	71.477	**QVMAX2**	11.465	**MT**
Worst	43.481	**MT**	45.054	**MT**	16.183	**CACLA1**

Table F.29 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 2.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	75.087	1.93073	0	0	24.913	1.93073
PF	42.111	3.7132	50.236	3.48368	7.654	0.36938
Max-Fair	8.971	0.88351	69.435	1.42716	21.594	0.64885
MT	36.491	4.9059	52.909	4.65631	10.6	0.56827
AS	11.806	1.10942	75.964	1.03303	12.23	0.19615
QV2	7.836	0.96858	77.024	1.09243	15.139	0.26748
QVMAX2	8.999	1.09014	78.36	1.09483	12.641	0.3914
ACLA	9.019	1.12784	78.129	1.14834	12.852	0.19393
CACLA1	9.438	1.27807	78.344	1.28475	12.218	0.21928
CACLA2	8.737	1.21485	78.379	1.20324	12.884	0.17113
Best	7.836	**QV2**	78.379	**CACLA2**	10.6	**MT**
Worst	36.491	**MT**	52.909	**MT**	15.139	**QV2**

Table F.30 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 2.75)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	75.812	1.04151	0	0	24.188	1.04151
PF	34.709	1.47262	58.07	1.40935	7.222	0.25956
Max-Fair	7.283	0.42878	73.818	0.53888	18.899	0.212
MT	24.935	4.00956	64.429	4.1159	10.636	0.31305
AS	9.969	0.91759	78.223	0.85177	11.808	0.1632
QV2	7.44	0.44293	81.119	0.39848	11.441	0.13699
QVMAX2	6.25	0.45378	81.101	0.5123	12.649	0.16493
ACLA	7.39	0.56826	81.247	0.51686	11.363	0.1727
CACLA1	7.556	0.66014	80.954	0.56418	11.49	0.18657
CACLA2	6.545	0.5458	81.707	0.4336	11.747	0.2285
Best	6.25	**QVMAX2**	81.707	**CACLA2**	10.636	**MT**
Worst	24.935	**MT**	64.429	**MT**	12.649	**QVMAX2**

Table F.31 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 3.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	77.03	2.01197	0	0	22.97	2.01197
PF	24.41	2.53372	67.774	2.39098	7.816	0.2135
Max-Fair	4.636	0.50441	76.781	0.48561	18.583	0.14388
MT	18.934	3.50238	68.068	3.1017	12.998	0.53802
AS	8.329	0.57981	79.435	0.55365	12.236	0.14703
QV2	6.608	0.45319	82.613	0.54258	10.78	0.26745
QVMAX2	7.911	0.91933	81.57	0.84851	10.518	0.17256
ACLA	6.054	0.38752	81.669	0.50973	12.278	0.3343
CACLA1	5.374	0.61235	82.746	0.55513	11.88	0.14592
CACLA2	4.926	0.53946	83.752	0.42091	11.322	0.18772
Best	4.926	**CACLA2**	83.752	**CACLA2**	10.518	**QVMAX2**
Worst	18.934	**MT**	68.068	**MT**	12.998	**MT**

Table F.32 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 3.25)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	77.316	2.34149	0	0.0003	22.684	2.34153
PF	19.121	2.9716	72.415	2.61034	8.464	0.59695
Max-Fair	3.861	0.39435	77.841	0.38883	18.298	0.27909
MT	16.676	1.47712	67.386	1.59207	15.938	0.91884
AS	8.279	0.69628	79.117	0.88309	12.604	0.28814
QV2	3.914	0.46027	83.329	0.34764	12.757	0.50341
QVMAX2	6.799	0.59144	82.411	0.46105	10.79	0.55072
ACLA	4.837	0.47413	83.361	0.32579	11.802	0.5273
CACLA1	5.362	0.75995	83.408	0.52626	11.231	0.61021
CACLA2	6.181	1.02909	83.219	0.86582	10.599	0.62814
Best	3.914	**QV2**	83.408	**CACLA1**	10.79	**QVMAX2**
Worst	16.676	**MT**	67.386	**MT**	15.938	**MT**

Table F.33 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 3.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	79.499	2.54721	0	0	20.501	2.54721
PF	14.287	2.5358	74.894	1.88515	10.82	0.8668
Max-Fair	3.173	0.30157	78.342	0.18569	18.485	0.34256
MT	18.089	1.11837	62.222	1.79001	19.689	0.89392
AS	9.156	1.07293	76.88	1.2746	13.964	0.4935
QV2	5.743	0.53586	81.934	0.20914	12.323	0.48408
QVMAX2	4.942	0.38008	82.57	0.27371	12.488	0.53992
ACLA	6.121	0.70424	82.287	0.38188	11.592	0.50263
CACLA1	5.746	1.08149	82.109	0.76822	12.145	0.83445
CACLA2	5.634	0.90354	83.182	0.53103	11.185	0.64171
Best	4.942	**QVMAX2**	83.182	**CACLA2**	11.185	**CACLA2**
Worst	18.089	**MT**	62.222	**MT**	19.689	**MT**

Table F.34 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 3.75)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	79.601	2.55257	0.002	0.00645	20.397	2.55345
PF	8.811	1.27174	74.353	0.87214	16.836	1.79399
Max-Fair	2.473	0.20083	77.364	0.43225	20.164	0.57138
MT	20.804	0.76013	55.39	1.2872	23.805	0.65569
AS	10.338	0.57765	73.487	0.88124	16.175	0.50267
QV2	3.262	0.293	80.385	1.21264	16.353	1.3818
QVMAX2	3.932	0.24705	80.405	0.86489	15.664	1.0041
ACLA	5.802	0.31541	80.008	0.75627	14.19	0.89181
CACLA1	3.486	0.4891	78.856	1.39085	17.658	1.74093
CACLA2	3.374	0.26744	80.558	0.59437	16.069	0.71366
Best	3.262	**QV2**	80.558	**CACLA2**	14.19	**ACLA**
Worst	20.804	**MT**	55.39	**MT**	23.805	**MT**

Table F.35 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 4.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	81.084	2.00524	0	0	18.916	2.00524
PF	7.053	0.76712	71.195	1.54196	21.752	1.92084
Max-Fair	2.249	0.11323	76.85	0.54672	20.902	0.63344
MT	22.547	0.99789	51.802	0.95868	25.651	0.45909
AS	11.736	0.48006	70.344	0.89206	17.92	0.475
QV2	14.716	0.49489	71.64	0.97779	13.644	0.62818
QVMAX2	7.04	0.33881	75.444	1.02639	17.516	0.77363
ACLA	2.971	0.21518	78.301	1.08465	18.728	1.19654
CACLA1	2.618	0.2491	77.299	1.40916	20.083	1.51552
CACLA2	2.84	0.29043	79.271	1.13259	17.889	1.30149
Best	2.618	**CACLA1**	79.271	**CACLA2**	13.644	**QV2**
Worst	22.547	**MT**	51.802	**MT**	25.651	**MT**

Table F.36 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 4.25)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	80.796	1.91221	0	0	19.205	1.91221
PF	6.212	0.61086	65.788	2.03086	28	2.58864
Max-Fair	2.176	0.10138	75.878	0.57691	21.946	0.62832
MT	23.996	0.93133	49.453	1.0345	26.551	0.43232
AS	14.108	1.17039	66.717	1.02101	19.176	0.49703
QV2	11.835	0.17972	72.205	1.51157	15.96	1.47608
QVMAX2	15.934	0.53139	64.972	1.31407	19.094	0.81526
ACLA	2.468	0.17984	74.438	1.52802	23.094	1.59385
CACLA1	2.447	0.20293	72.905	2.29759	24.648	2.39229
CACLA2	3.078	0.25672	77.266	1.27282	19.655	1.37885
Best	2.447	**CACLA1**	72.905	**CACLA1**	15.96	**QV2**
Worst	23.996	**MT**	49.453	**MT**	26.551	**MT**

Table F.37 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 4.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$.

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	82.765	1.49012	0	0	17.235	1.49012
PF	4.841	0.41783	59.058	1.66854	36.101	1.93081
Max-Fair	1.985	0.07451	73.188	0.91595	24.828	0.96932
MT	25.371	0.7904	47.119	0.6501	27.51	0.33354
AS	15.253	0.95131	64.259	0.91632	20.489	0.39975
QV2	18.815	0.3974	64.045	0.87302	17.14	0.55748
QVMAX2	14.099	0.78929	64.891	1.11373	21.01	0.54846
ACLA	2.672	0.12905	69.072	1.32027	28.255	1.36945
CACLA1	8.553	0.23713	61.041	1.13343	30.406	1.24953
CACLA2	2.209	0.11315	68.035	1.5519	29.756	1.63644
Best	2.209	**CACLA2**	69.072	**ACLA**	17.14	**QV2**
Worst	25.371	**MT**	47.119	**MT**	30.406	**CACLA1**

Table F.38 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 4.75)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	82.723	1.549	0	0	17.277	1.549
PF	4.344	0.39986	52.853	3.33092	42.802	3.67692
Max-Fair	1.93	0.07657	70.868	1.62455	27.202	1.68405
MT	25.879	0.6548	45.729	0.82912	28.393	0.52552
AS	16.431	0.73163	62.218	0.75386	21.351	0.26715
QV2	7.791	0.75277	63.003	2.78222	29.206	2.04565
QVMAX2	16.108	0.59677	59.639	2.17392	24.253	1.60258
ACLA	6.652	0.39293	61.867	2.65229	31.481	2.29162
CACLA1	2.071	0.08945	60.19	4.11122	37.739	4.18493
CACLA2	2.528	0.13091	80.92	0.30227	16.552	0.4209
Best	2.071	**CACLA1**	80.92	**CACLA2**	16.552	**CACLA2**
Worst	25.879	**MT**	45.729	**MT**	37.739	**CACLA1**

Table F.39 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 5.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	84.231	1.34479	0	0	15.769	1.34479
PF	3.587	0.29354	46.972	2.9142	49.441	3.10672
Max-Fair	1.884	0.05212	67.268	1.69273	30.848	1.73932
MT	27.319	0.84226	44.015	0.63881	28.666	0.49225
AS	17.35	1.05177	60.645	1.13205	22.005	0.28616
QV2	2.22	0.23121	54.64	3.54375	43.14	3.55046
QVMAX2	26.162	0.4125	53.479	0.62465	20.359	0.31086
ACLA	9.91	0.47095	55.153	1.84256	34.937	1.44023
CACLA1	8.102	0.25751	50.552	1.96413	41.346	2.15197
CACLA2	1.79	0.05448	70.77	1.83284	27.44	1.8718
Best	1.79	**CACLA2**	70.77	**CACLA2**	20.359	**QVMAX2**
Worst	27.319	**MT**	44.015	**MT**	43.14	**QV2**

Table F.40 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 5.25)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	83.932	0.90401	0	0	16.068	0.90401
PF	3.337	0.2515	40.073	1.94278	56.59	2.10918
Max-Fair	1.824	0.03056	62.694	1.5894	35.482	1.61417
MT	27.722	0.29454	42.748	0.46109	29.53	0.40205
AS	18.756	1.87733	58.94	1.48403	22.304	0.52154
QV2	18.113	0.34147	51.703	0.64023	30.183	0.50177
QVMAX2	13.876	0.75269	49.309	1.73329	36.815	1.05454
ACLA	8.852	0.47764	53.895	1.62114	37.253	1.18876
CACLA1	6.373	0.21676	43.508	1.51071	50.119	1.6974
CACLA2	11.887	0.16354	56.626	0.65347	31.487	0.65319
Best	6.373	**CACLA1**	56.626	**CACLA2**	50.119	**CACLA1**
Worst	27.722	**MT**	42.748	**MT**	22.304	**AS**

Table F.41 Percentages of TTIs (Mean and STD) in the exploitation stage when $\mathcal{S}_t^C \in \{\mathcal{UFF}, \mathcal{FAF}, \mathcal{OFF}\}$ for NGMN Fairness Requirement based on NAUT-MMF with static windowing factor: $(\rho = 5.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Mean / STD / RL Alg.	Mean[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{UFF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{FAF}$	Mean[%] $\mathcal{S}_t^C \in \mathcal{OFF}$	STD[%] $\mathcal{S}_t^C \in \mathcal{OFF}$
Max-Th	83.299	1.28587	0	0	16.701	1.28587
PF	3.175	0.20695	34.778	1.98904	62.047	2.13857
Max-Fair	1.779	0.05277	56.929	2.94159	41.292	2.98367
MT	28.818	0.56325	40.885	1.00522	30.297	0.54536
AS	19.263	1.39627	57.961	1.10097	22.776	0.37145
QV2	7.927	7.78921	38.52	4.76863	53.553	12.35889
QVMAX2	23.026	1.7647	52.789	1.92558	24.185	3.55169
ACLA	16.284	6.23825	47.955	2.25826	35.761	8.34952
CACLA1	10.721	2.28147	42.78	1.47608	46.499	3.669
CACLA2	13.316	5.71871	53.984	6.65361	32.699	12.28351
Best	7.927	**QV2**	57.961	**AS**	22.776	**AS**
Worst	28.818	**MT**	38.52	**QV2**	53.553	**QV2**

Table F.42 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 2.0)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	7.594	0.30057	58.073	1.80422	34.333	1.52894
QVMAX2	7.464	0.45808	60.083	2.02339	32.453	1.59832
ACLA	14.62	0.5198	56.673	2.09446	28.707	1.69233
CACLA1	13.973	0.4039	59.133	1.88705	26.894	1.51157
CACLA2	19.699	0.53791	59.9	1.58685	20.401	1.13742
Best	7.464	**QVMAX2**	60.083	**QVMAX2**	20.401	**CACLA2**
Worst	19.699	**CACLA2**	58.073	**QV2**	34.333	**QV2**

Table F.43 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 2.25)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	8.283	0.46784	70.904	1.23141	20.814	0.82228
QVMAX2	7.359	0.48422	71.472	1.18441	21.169	0.78002
ACLA	8.562	0.32911	68.061	1.41391	23.377	1.10802
CACLA1	15.578	0.59426	70.493	1.19739	13.929	0.63806
CACLA2	22.566	1.0229	70.185	1.53633	7.249	0.53006
Best	7.359	**QVMAX2**	71.472	**QVMAX2**	7.249	**CACLA2**
Worst	15.578	**CACLA1**	68.061	**ACLA**	23.377	**ACLA**

Table F.44 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 2.5)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	6.693	0.82605	77.018	1.09324	16.288	0.32208
QVMAX2	4.792	0.4513	78.352	1.09502	16.856	0.72986
ACLA	7.299	0.52073	78.123	1.14908	14.578	0.67777
CACLA1	10.156	0.46929	78.338	1.28535	11.506	0.83023
CACLA2	12.004	0.45227	78.373	1.20307	9.623	0.77397
Best	6.693	**QV2**	78.373	**CACLA2**	9.623	**CACLA2**
Worst	12.004	**CACLA2**	77.018	**QV2**	16.856	**QVMAX2**

Table F.45 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 2.75)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	4.185	0.21366	81.113	0.39749	14.702	0.205
QVMAX2	3.636	0.25557	81.095	0.51223	15.269	0.29148
ACLA	8.396	0.25703	81.241	0.51645	10.363	0.56632
CACLA1	8.695	0.19864	80.948	0.564	10.357	0.38
CACLA2	11.934	0.24724	81.7	0.43413	6.366	0.28494
Best	3.636	**QVMAX2**	81.7	**CACLA2**	6.366	**CACLA2**
Worst	11.934	**CACLA2**	80.948	**CACLA1**	15.269	**QVMAX2**

Table F.46 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 3.0)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	3.487	0.3554	82.605	0.54373	13.908	0.26468
QVMAX2	2.971	0.25635	81.563	0.84891	15.466	0.61894
ACLA	5.438	0.39658	81.663	0.5094	12.899	0.18009
CACLA1	8.292	0.26807	82.737	0.55594	8.971	0.2944
CACLA2	10.336	0.212	83.745	0.42098	5.919	0.2325
Best	2.971	**QVMAX2**	83.745	**CACLA2**	5.919	**CACLA2**
Worst	10.336	**CACLA2**	81.563	**QVMAX2**	15.466	**QVMAX2**

Table F.47 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 3.25)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	1.517	0.16563	83.322	0.34776	15.161	0.36053
QVMAX2	3.791	0.36223	82.403	0.46065	13.806	0.30307
ACLA	4.784	0.08729	83.355	0.32567	11.861	0.26931
CACLA1	8.162	0.2067	83.4	0.52729	8.438	0.34071
CACLA2	10.014	0.45478	83.21	0.86673	6.776	0.48516
Best	1.517	**QV2**	83.4	**CACLA1**	6.776	**CACLA2**
Worst	10.014	**CACLA2**	82.403	**QVMAX2**	15.161	**QV2**

Table F.48 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 3.5)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	8.488	0.21813	81.927	0.20944	9.585	0.23742
QVMAX2	3.782	0.24319	82.563	0.2728	13.654	0.28957
ACLA	5.07	0.16124	82.28	0.38251	12.65	0.28026
CACLA1	7.784	0.37729	82.099	0.76837	10.116	0.46623
CACLA2	10.752	0.28525	83.174	0.53152	6.074	0.35723
Best	3.782	**QVMAX2**	83.174	**CACLA2**	6.074	**CACLA2**
Worst	10.752	**CACLA2**	81.927	**QV2**	13.654	**QVMAX2**

Table F.49 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 3.75)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	1.986	0.27936	80.379	1.21221	17.636	1.17265
QVMAX2	2.997	0.20787	80.399	0.86458	16.604	0.8371
ACLA	5.183	0.1874	80.002	0.75633	14.816	0.59333
CACLA1	8.849	0.54991	78.847	1.39176	12.304	0.84683
CACLA2	13.201	0.49586	80.552	0.5937	6.247	0.12267
Best	1.986	**QV2**	80.552	**CACLA2**	6.247	**CACLA2**
Worst	13.201	**CACLA2**	78.847	**CACLA1**	17.636	**QV2**

Table F.50 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 4.0)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	5.518	0.30461	71.634	0.97712	22.848	0.80028
QVMAX2	3.689	0.28448	75.437	1.02574	20.874	0.9874
ACLA	11.261	0.93905	78.296	1.0835	10.443	0.24229
CACLA1	10.713	0.63853	77.291	1.40856	11.996	0.79677
CACLA2	14.676	0.86677	79.264	1.1323	6.06	0.29039
Best	3.689	**QVMAX2**	79.264	**CACLA2**	6.06	**CACLA2**
Worst	14.676	**CACLA2**	71.634	**QV2**	22.848	**QV2**

Table F.51 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 4.25)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	4.224	0.09428	72.199	1.51096	23.577	1.4338
QVMAX2	5.533	0.18062	64.967	1.31353	29.5	1.26313
ACLA	5.285	0.145	74.434	1.52652	20.281	1.41885
CACLA1	9.902	0.73053	72.899	2.29728	17.199	1.58247
CACLA2	14.552	1.10224	77.26	1.27237	8.189	0.23279
Best	4.224	**QV2**	77.26	**CACLA2**	8.189	**CACLA2**
Worst	14.552	**CACLA2**	64.967	**QVMAX2**	29.5	**QVMAX2**

Table F.52 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 4.5)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	5.385	0.35845	64.04	0.87299	30.575	0.7499
QVMAX2	5.488	0.34711	64.887	1.11393	29.626	1.15308
ACLA	5.615	0.20021	69.069	1.32086	25.316	1.24015
CACLA1	11.23	0.35691	61.036	1.13358	27.734	0.80165
CACLA2	26.138	1.35105	68.03	1.55239	5.833	0.23439
Best	5.385	**QV2**	69.069	**ACLA**	5.833	**CACLA2**
Worst	26.138	**CACLA2**	61.036	**CACLA1**	30.575	**QV2**

Table F.53 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 4.75)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	8.219	0.27052	62.999	2.78175	28.782	2.69807
QVMAX2	4.817	0.23063	59.636	2.17315	35.548	2.22795
ACLA	6.759	0.25818	61.864	2.65239	31.377	2.52827
CACLA1	13.543	1.52251	60.186	4.11129	26.271	2.60103
CACLA2	12.912	0.25056	80.915	0.30205	6.173	0.07093
Best	4.817	**QVMAX2**	80.915	**CACLA2**	6.173	**CACLA2**
Worst	13.543	**CACLA1**	59.636	**QVMAX2**	35.548	**QVMAX2**

Table F.54 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 5.0)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	9.362	0.25771	54.638	3.54342	36.001	3.66298
QVMAX2	4.725	0.24535	53.474	0.6251	41.801	0.65302
ACLA	5.593	0.24343	55.15	1.84307	39.257	1.76926
CACLA1	13.116	0.66407	50.548	1.96446	36.336	1.32733
CACLA2	22.198	1.41836	70.767	1.83226	7.035	0.45628
Best	4.725	**QVMAX2**	70.767	**CACLA2**	7.035	**CACLA2**
Worst	22.198	**CACLA2**	50.548	**CACLA1**	41.801	**QVMAX2**

Table F.55 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 5.25)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	8.662	0.21135	51.699	0.64014	39.639	0.55021
QVMAX2	6.21	0.24924	49.306	1.73255	44.484	1.56608
ACLA	9.279	0.13921	53.891	1.62096	36.83	1.54538
CACLA1	15.556	0.65451	43.503	1.51038	40.94	0.90537
CACLA2	20.772	0.33789	56.621	0.65425	22.607	0.36745
Best	6.21	**QVMAX2**	56.621	**CACLA2**	22.607	**CACLA2**
Worst	20.772	**CACLA2**	43.503	**CACLA1**	44.484	**QVMAX2**

Table F.56 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^F = \{\pm 1\}$ and $\mathcal{RW}_t^F \in \mathbb{R}_{(-1,1)}$. The NGMN Fairness Requirement is based on NAUT-MMF user rates with static windowing factor: $(\rho = 5.5)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^F = -1$	Punish Reward STD[%] $\mathcal{RW}_t^F = -1$	Max. Reward Mean[%] $\mathcal{RW}_t^F = 1$	Max. Reward STD[%] $\mathcal{RW}_t^F = 1$	Moderate Reward Mean[%] $\mathcal{RW}_t^F \neq \pm 1$	Moderate Reward STD[%] $\mathcal{RW}_t^F \neq \pm 1$
QV2	4.306	0.92454	38.518	4.76842	57.176	5.64068
QVMAX2	5.075	0.51652	52.785	1.92479	42.14	1.47743
ACLA	6.484	0.11593	47.95	2.2572	45.566	2.35372
CACLA1	14.039	3.53231	42.775	1.47493	43.186	2.25868
CACLA2	26.466	11.77444	53.98	6.6523	19.554	5.22598
Best	4.306	**QV2**	53.98	**CACLA2**	19.554	**CACLA2**
Worst	26.466	**CACLA2**	38.518	**QV2**	57.176	**QV2**

F.4 Summary

For the DSR-SMOO problems with AUT-EMF observations, CACLA2 shows the best performance from the viewpoint of the mean percentage of feasible TTIs $\overline{p}_{TTI}^{F,FAF}$ in almost all the cases of the CQI aggregation schemes, by minimizing at the same time, the STD values when compared with other classical approaches such as AS and MT. Also, the CACLA2 policies provide the lowest mean percentage of TTIs $\overline{p}_{TTI}^{F,OFF}$ when the scheduler stays over-fair for the following CQI aggregation schemes: $\{Top4, N_{CT}=128\}$, $\{Top4, N_{CT}=256\}$, $\{Top4, N_{CT}=512\}$, $\{Top5, N_{CT}=128\}$ and $\{Top5, N_{CT}=512\}$. When the mean percentage of TTIs with unfair states $\overline{p}_{TTI}^{F,UFF}$ is considered, CACLA2 policy is the best option when four CQI aggregation techniques are applied such as: $\{Top3, N_{CT}=512\}$, $\{Top4, N_{CT}=64\}$, $\{Top5, N_{CT}=64\}$ and $\{Top5, N_{CT}=128\}$. Excepting the cases of aggregation schemes $\{Top3, N_{CT}=256\}$, $\{Top4, N_{CT}=64\}$ and $\{Top5, N_{CT}=64\}$, CACLA2 policies indicate the best performance from the viewpoint of the percentage of TTIs with moderate rewards. To conclude, CACLA2 scheduling policies offer the best sustainability by maximizing the percentages of feasible TTIs when the AUT-EMF observations are used.

In the case of the DSR-SMOO problems which make use of AUT-MMF observations, the CACLA2 policy provides the best performance from the viewpoint of $\overline{p}_{TTI}^{F,FAF}$ when compared with any other candidate if the windowing factor belongs to $\rho \in [2.5; 4.0]$. In these cases, the MT methodology provides the lowest percentage of $\overline{p}_{TTI}^{F,FAF}$ when compared against other policies. For the entire considered range of windowing factors, CACLA2 provides the lowest mean percentage of moderate rewards $\overline{p}_{TTI}^{F,mRW}$. This advantage is not fully exploited since CACLA2 policies provide the highest amount of TTIs when the testing rewards are punishments. However, CACLA2 policies are sustainable for the aforementioned interval of $\rho \in [2.5; 4.0]$ when $\overline{p}_{TTI}^{F,FAF}$ is maximized for each case.

Appendix G

Performance Evaluation of Sustainable Scheduling Policies Focusing on GBR Requirement

G.1 Appendix Outline

The performance evaluation of scheduling policies being focused on the GBR objective for the infinite buffer, CBR and VBR traffic types are analysed in this section. Basically, the simulation results presented in this section extend the performances of the sustainable scheduling policies highlighted in Sub-section 6.3.4 from Chapter 6. The experimental results are conducted through two directions such as: mean percentages of TTIs for the GBR satisfaction levels and the mean percentage of TTIs when the scheduler rewards for the GBR objective (see Sub-section 6.3.3) are maximized, moderate or punishment. If all active bearers are 100% satisfied from the viewpoint of the GBR objective, then the feasible state is reached. A crucial role in the GBR objective satisfaction is played by the windowing factor which is used to compute the AUT-MMF observations. The scheduling policies are trained based on multiple windowing factors in order to find the optimum range for the simulation scenario exposed in Table 6.3.

G.2 Percentages of TTIs for the GBR User Satisfaction Levels Based on Infinite Buffer, CBR and VBR Traffic Types

The percentages of TTIs with the GBR satisfaction levels from 91% to 100% are presented for infinite buffer, CBR and VBR traffic types. The scheduling policies are trained based on the following RL approaches: QV, QV2, QVMAX, QVMAX2 and ACLA approaches. The obtained sets of scheduling policies are compared in the exploitation stage against the simple scheduling rules such as: GPF-BF, GPF-mM, GPF-RAD and GPF-LM. The sustainable policies are trained and evaluated based on different static windowing factors belonging to $\rho \in [2.0, 5.5]$ with a step of 0.5 in order to find the optimal range in which the mean percentage of feasible TTIs $\overline{p}_{TTI}^{G,100\%}$ is maximized. It is expected that if the windowing factor is too small, then the DSR-SMOO MDP problems are non-episodic and implicitly, the feasible state when all active bearers are 100% satisfied will not be reached. On the other side, if the windowing factor is too large, then the mean percentage of feasible TTIs can increase substantially but, the controller cannot detect the real benefits of applying certain rule in different state leading in this way, to un-optimal scheduling decisions at each TTI. This situation can be detected in two ways: based on the standard deviation values or based on the testing rewards. If the STD of the mean percentage of TTIs is too large, then the scheduling policy is unsuitable to be applied in real practice. Also, if the number of moderate or punishment rewards is very high when compared with the maximum rewards, then the windowing factor is not optimal. The same CQI aggregation technique is considered in the controller state space computation for all simulation results in terms of $(Top3, N_{CT} = 64)$. The rest of this sub-section is organized as follows: Tables G.1 to G.8 highlights the performance of scheduling policies for the infinite buffer traffic type, Tables G.9 to G.16 evaluates the obtained policies for the CBR traffic type and finally, Tables G.17 to G.24 presents the advantages of using the obtained policies against the existing techniques for the VBR traffic type.

Table G.1 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 2.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0	0.00032	0.108	0.018	0.142	0.12451	20.969	2.16945
92% Sat.	0	0.00032	0.034	0.01187	0.139	0.12471	17.481	2.17617
93% Sat.	0	0.00032	0.023	0.007	0.137	0.12446	13.987	1.76643
94% Sat.	0	0.00032	0.01	0.00352	0.135	0.12529	10.877	1.63496
95% Sat.	0	0.00032	0.006	0.00234	0.135	0.12529	8.983	1.43552
96% Sat.	0	0.00032	0.004	0.00197	0.135	0.12529	6.72	1.29418
97% Sat.	0	0.00032	0.004	0.00184	0.135	0.12529	4.486	0.92808
98% Sat.	0	0.00032	0.004	0.00184	0.135	0.12529	3.159	0.71575
99% Sat.	0	0.00032	0.004	0.00184	0.135	0.12529	2.316	0.59207
100% Sat.	0	0.00032	0.004	0.00184	0.135	0.12529	1.633	0.48823

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	20.988	2.0845	0.317	0.23814	0.001	0.00185	20.544	1.99548	0.107	0.02287
92% Sat.	17.495	2.11	0.184	0.13743	0	0.0007	17.078	1.96323	0.037	0.01004
93% Sat.	13.967	1.7321	0.164	0.13835	0	0.0007	13.623	1.60104	0.022	0.00626
94% Sat.	10.887	1.58265	0.132	0.12758	0	0.0007	10.565	1.48164	0.009	0.0022
95% Sat.	9.002	1.36391	0.076	0.06233	0	0.0007	8.718	1.30183	0.006	0.00164
96% Sat.	6.723	1.24085	0.064	0.05776	0	0.0007	6.522	1.21116	0.004	0.00137
97% Sat.	4.534	0.91745	0.031	0.02188	0	0.0007	4.357	0.87145	0.003	0.00125
98% Sat.	3.199	0.68914	0.028	0.02211	0	0.0007	3.068	0.67443	0.003	0.00105
99% Sat.	2.331	0.59307	0.028	0.02211	0	0.0007	2.247	0.56993	0.003	0.00105
100% Sat.	1.668	0.4892	0.02	0.02006	0	0.0007	1.61	0.48199	0.003	0.00105

Table G.2 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 2.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.002	0.00103	1.117	0.12545	0.974	0.12659	41.701	1.39514
92% Sat.	0.002	0.00084	0.577	0.0807	0.875	0.11355	39.137	1.48852
93% Sat.	0.002	0.00084	0.353	0.04955	0.731	0.13274	34.465	1.61867
94% Sat.	0.002	0.00084	0.151	0.02511	0.61	0.12043	29.703	1.68504
95% Sat.	0.002	0.00084	0.081	0.02093	0.61	0.12043	26.926	1.75427
96% Sat.	0.002	0.00084	0.043	0.00946	0.61	0.12043	22.112	1.74474
97% Sat.	0.002	0.00084	0.027	0.00666	0.61	0.12043	17.741	1.48122
98% Sat.	0.002	0.00084	0.024	0.00598	0.61	0.12043	14.023	1.31848
99% Sat.	0.002	0.00084	0.023	0.00657	0.61	0.12043	11.335	1.07553
100% Sat.	0.002	0.00084	0.023	0.00643	0.61	0.12043	7.953	0.76684

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	38.884	1.54848	0.033	0.01299	41.698	1.45723	1.477	0.30219	39.483	1.84398
92% Sat.	36.376	1.59703	0.029	0.00981	39.173	1.57286	1.439	0.29788	36.98	1.96192
93% Sat.	32.086	1.75802	0.022	0.0083	34.545	1.69911	0.879	0.18227	32.409	1.89134
94% Sat.	27.917	1.73947	0.022	0.0083	29.793	1.73275	0.709	0.13323	27.726	1.88563
95% Sat.	25.176	1.81169	0.022	0.0083	27.014	1.82702	0.709	0.13323	24.993	1.93808
96% Sat.	20.528	1.805	0.022	0.0083	22.186	1.80981	0.605	0.12196	20.482	1.69557
97% Sat.	16.255	1.50284	0.022	0.0083	17.794	1.48555	0.604	0.12366	16.327	1.4714
98% Sat.	12.646	1.33381	0.022	0.0083	14.031	1.33127	0.594	0.13032	12.803	1.24798
99% Sat.	10.016	1.0848	0.022	0.0083	11.359	1.07644	0.594	0.13032	10.473	0.99171
100% Sat.	6.712	0.79949	0.022	0.0083	7.985	0.76942	0.594	0.13047	7.404	0.7403

Table G.3 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 3.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.028	0.0249	2.835	0.32995	4.129	0.30642	52.016	1.79243
92% Sat.	0.028	0.0249	1.671	0.25836	4.011	0.29316	50.634	2.01023
93% Sat.	0.028	0.0249	1.045	0.2086	3.913	0.28041	46.799	2.2157
94% Sat.	0.028	0.02493	0.498	0.09887	3.562	0.28081	43.039	2.40098
95% Sat.	0.028	0.02493	0.351	0.08144	3.562	0.28081	40.847	2.50948
96% Sat.	0.028	0.02493	0.179	0.0426	3.562	0.28081	36.307	2.88941
97% Sat.	0.028	0.02493	0.119	0.03261	3.562	0.28081	32.029	2.79919
98% Sat.	0.028	0.02493	0.104	0.03347	3.558	0.28037	26.924	2.47758
99% Sat.	0.028	0.02493	0.099	0.03222	3.558	0.28037	23.717	2.42379
100% Sat.	0.028	0.02493	0.098	0.0323	3.558	0.28037	18.196	1.76855

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	13.701	0.93212	10.539	0.91606	2.786	0.32975	51.229	1.75495	51.917	1.79673
92% Sat.	11.7	0.8668	8.945	0.88927	1.649	0.27606	49.672	1.94162	50.544	2.00981
93% Sat.	9.827	0.77223	7.421	0.78932	1.06	0.19406	45.696	2.10097	46.755	2.1411
94% Sat.	8.079	0.7039	6.006	0.68179	0.496	0.09915	41.723	2.24866	43.003	2.34444
95% Sat.	7.629	0.71895	5.694	0.67359	0.359	0.07748	39.356	2.28076	40.786	2.41018
96% Sat.	6.588	0.70242	4.799	0.63591	0.182	0.03767	34.676	2.60231	36.265	2.78802
97% Sat.	5.91	0.6656	4.322	0.61967	0.116	0.02931	30.388	2.48085	31.979	2.65968
98% Sat.	5.505	0.63881	3.941	0.60214	0.103	0.02933	25.447	2.11959	26.861	2.34955
99% Sat.	5.301	0.65089	3.865	0.60147	0.099	0.02848	22.329	2.08492	23.644	2.30311
100% Sat.	5.231	0.64001	3.849	0.59844	0.097	0.02781	17.368	1.55059	18.118	1.70531

Table G.4 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 3.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.09	0.05733	5.04	0.69418	4.501	0.44934	55.892	0.96614
92% Sat.	0.089	0.05643	3.154	0.46353	4.301	0.45603	55.188	1.03225
93% Sat.	0.036	0.02047	1.876	0.27927	4.076	0.44352	52.522	1.30218
94% Sat.	0.035	0.02047	1.029	0.17104	3.714	0.45697	49.845	1.4457
95% Sat.	0.035	0.02047	0.707	0.13161	3.714	0.45697	48.467	1.75965
96% Sat.	0.035	0.02047	0.39	0.069	3.714	0.45697	44.596	2.15706
97% Sat.	0.035	0.02047	0.224	0.03699	3.714	0.45697	40.746	2.67882
98% Sat.	0.035	0.02047	0.169	0.03015	3.668	0.45377	35.725	2.92902
99% Sat.	0.035	0.02047	0.151	0.02802	3.668	0.45377	32.363	3.18143
100% Sat.	0.035	0.02047	0.146	0.02757	3.668	0.45377	25.037	3.02185

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	17.308	1.89144	4.986	0.74642	5.089	0.75599	0.705	0.22268	18.721	0.69659
92% Sat.	15.922	1.88576	3.075	0.51967	3.225	0.55009	0.553	0.19317	15.002	0.73016
93% Sat.	13.927	1.80319	1.847	0.33892	1.938	0.35565	0.228	0.06703	11.869	0.62382
94% Sat.	12.76	1.76886	1.008	0.22415	1.051	0.21691	0.172	0.04651	8.69	0.6105
95% Sat.	12.335	1.72771	0.685	0.15613	0.714	0.14541	0.172	0.04651	7.081	0.55496
96% Sat.	11.24	1.62658	0.388	0.09599	0.401	0.07973	0.141	0.02922	5.265	0.48945
97% Sat.	10.06	1.71584	0.215	0.04302	0.231	0.03564	0.136	0.03146	3.897	0.44992
98% Sat.	8.731	1.86291	0.166	0.03064	0.18	0.03118	0.13	0.02577	3.166	0.45384
99% Sat.	7.687	1.62642	0.149	0.02672	0.159	0.02525	0.13	0.02577	2.85	0.4551
100% Sat.	7.27	1.51744	0.145	0.02402	0.154	0.02381	0.13	0.02577	2.696	0.45087

Table G.5 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 4.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.142	0.06831	6.876	0.34475	4.972	0.24847	58.917	0.45892
92% Sat.	0.129	0.06395	4.542	0.29992	4.764	0.27274	58.489	0.51086
93% Sat.	0.092	0.0402	2.963	0.17193	4.517	0.27993	56.204	0.73032
94% Sat.	0.054	0.01974	1.636	0.11498	4.333	0.31137	54.53	0.8003
95% Sat.	0.054	0.01974	1.223	0.10118	4.333	0.31137	53.765	0.87848
96% Sat.	0.054	0.01974	0.716	0.05666	4.333	0.31137	51.148	1.02487
97% Sat.	0.053	0.01969	0.387	0.03555	4.333	0.31137	48.72	1.2237
98% Sat.	0.053	0.01969	0.25	0.02411	4.196	0.30335	44.512	1.81407
99% Sat.	0.053	0.01969	0.208	0.02384	4.196	0.30335	42.024	2.07234
100% Sat.	0.053	0.01969	0.201	0.02292	4.196	0.30335	34.312	2.24931

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	33.603	0.75841	14.971	0.79979	58.361	0.51268	23.824	0.95178	57.143	0.48033
92% Sat.	30.232	0.80281	13.994	0.80509	57.951	0.5628	21.865	0.96225	56.704	0.48223
93% Sat.	25.897	0.79661	12.778	0.76571	55.633	0.75199	18.92	0.99051	54.299	0.66857
94% Sat.	21.605	0.73556	11.88	0.71117	53.934	0.83052	16.347	0.919	52.571	0.7292
95% Sat.	19.411	0.71109	11.813	0.70813	53.186	0.91971	15.445	0.88049	51.774	0.74793
96% Sat.	15.69	0.58626	10.899	0.62034	50.507	1.08119	13.583	0.80354	49.088	0.93927
97% Sat.	12.922	0.55348	10.103	0.5977	48.066	1.31155	11.992	0.81204	46.59	1.1287
98% Sat.	10.819	0.4821	9.613	0.5418	43.799	1.80171	10.331	0.7276	42.306	1.64281
99% Sat.	9.407	0.41984	9.574	0.53738	41.298	2.0208	9.11	0.61246	39.826	1.84739
100% Sat.	8.537	0.41851	9.559	0.53881	33.559	2.21081	7.948	0.5568	32.161	2.06982

Table G.6 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 4.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.875	0.26719	9.494	0.43242	5.524	0.33465	60.525	0.42855
92% Sat.	0.675	0.16334	6.591	0.38562	5.431	0.31332	60.1	0.40544
93% Sat.	0.148	0.06336	4.157	0.34063	5.295	0.30992	58.428	0.56941
94% Sat.	0.114	0.04331	2.78	0.29412	5.14	0.33509	56.993	0.69658
95% Sat.	0.114	0.04331	2.276	0.26303	5.14	0.33498	56.467	0.76636
96% Sat.	0.1	0.03935	1.23	0.12818	5.14	0.33498	54.417	0.77581
97% Sat.	0.1	0.03971	0.736	0.06354	5.14	0.33498	52.172	1.14991
98% Sat.	0.1	0.03971	0.405	0.05007	4.953	0.32881	48.666	1.19641
99% Sat.	0.1	0.03971	0.301	0.04574	4.953	0.32881	46.604	1.39871
100% Sat.	0.1	0.03971	0.289	0.04812	4.953	0.32881	38.887	1.60578

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	8.065	0.3772	9.543	0.45256	28.78	5.34676	9.83	0.54858	60.523	0.46299
92% Sat.	5.72	0.2982	6.699	0.35232	28.552	5.35089	6.903	0.37786	60.103	0.46894
93% Sat.	3.625	0.29714	4.195	0.296	27.331	5.13141	4.371	0.36847	58.419	0.62482
94% Sat.	2.49	0.26111	2.78	0.23214	26.35	5.02486	2.977	0.26775	56.973	0.74889
95% Sat.	2.168	0.23346	2.255	0.187	25.932	4.96599	2.426	0.24225	56.421	0.81005
96% Sat.	1.482	0.24499	1.271	0.11758	25.034	4.67252	1.369	0.1571	54.386	0.79707
97% Sat.	1.201	0.2447	0.763	0.09518	23.733	4.59954	0.902	0.07318	52.133	1.2209
98% Sat.	0.976	0.24802	0.413	0.05111	22.183	4.34916	0.486	0.03348	48.599	1.28319
99% Sat.	0.925	0.24361	0.319	0.0423	21.098	4.27587	0.389	0.04288	46.564	1.46459
100% Sat.	0.918	0.24579	0.301	0.04381	18.869	3.44817	0.322	0.04936	38.87	1.59298

Table G.7 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 5.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	2.557	0.54782	13.376	0.62152	6.343	0.31564	62.428	0.71903
92% Sat.	2.285	0.44416	10.271	0.62725	6.266	0.31223	62.057	0.74296
93% Sat.	1.235	0.3277	6.956	0.4804	6.149	0.32212	60.808	0.74552
94% Sat.	0.855	0.32216	4.716	0.43317	6.105	0.32973	59.828	0.67904
95% Sat.	0.855	0.32216	4.201	0.40107	6.105	0.32973	59.274	0.76821
96% Sat.	0.561	0.22693	2.513	0.29074	6.105	0.32973	57.791	0.91464
97% Sat.	0.537	0.19674	1.738	0.27033	6.105	0.32973	56.194	1.04531
98% Sat.	0.534	0.19416	0.853	0.2075	5.95	0.31138	53.329	1.35297
99% Sat.	0.534	0.19416	0.716	0.20997	5.95	0.31138	51.406	1.41254
100% Sat.	0.534	0.19416	0.702	0.20667	5.95	0.31138	44.163	1.90356

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	56.257	1.21406	20.81	0.42946	9.971	0.58063	29.515	1.83771	45.761	1.74041
92% Sat.	55.409	1.2151	17.959	0.37596	7.64	0.48098	28.19	1.73234	42.113	1.80924
93% Sat.	53.621	1.26149	14.478	0.36885	4.997	0.39304	25.208	1.68143	37.473	1.76264
94% Sat.	51.435	1.33745	11.623	0.37515	3.391	0.36803	22.867	1.58713	32.207	1.69696
95% Sat.	50.035	1.38033	10.544	0.36769	3.235	0.36102	22.169	1.52792	29.143	1.54257
96% Sat.	47.16	1.53605	8.058	0.33983	1.881	0.21121	20.017	1.43582	24.707	1.37156
97% Sat.	43.858	1.78277	6.54	0.34302	1.287	0.20893	18.824	1.28734	21.117	1.31116
98% Sat.	39.718	1.75438	4.889	0.33577	0.701	0.20063	16.84	1.16891	18.407	1.09354
99% Sat.	36.611	1.70702	3.973	0.31401	0.667	0.19651	15.422	1.01402	16.452	1.04766
100% Sat.	31.228	1.31711	3.512	0.32191	0.665	0.19641	14.051	0.84988	15.49	0.93369

Table G.8 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 5.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer.

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	4.586	0.7028	15.617	0.8655	6.267	0.16652	62.941	0.37657
92% Sat.	3.298	0.5791	12.101	0.73863	6.219	0.15535	62.574	0.36266
93% Sat.	2.389	0.46116	8.549	0.66222	6.151	0.15782	61.27	0.37293
94% Sat.	1.035	0.27766	5.671	0.50732	6.129	0.15464	60.373	0.40694
95% Sat.	1.035	0.27766	5.061	0.4614	6.129	0.15464	59.879	0.50701
96% Sat.	0.652	0.17378	3.153	0.27822	6.129	0.15464	58.515	0.49117
97% Sat.	0.64	0.17047	2.348	0.2169	6.129	0.15464	57.265	0.59558
98% Sat.	0.601	0.15209	0.939	0.1604	5.981	0.15009	54.673	0.8844
99% Sat.	0.601	0.15209	0.723	0.17442	5.981	0.15009	53.216	1.0612
100% Sat.	0.601	0.15209	0.697	0.17698	5.981	0.15009	46.759	1.54367

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	24.843	0.6427	34.034	0.84436	6.614	0.62967	58.852	0.34496	62.962	0.31873
92% Sat.	22.637	0.6441	30.984	0.79296	5.907	0.63755	58.07	0.34341	62.602	0.30565
93% Sat.	19.058	0.63404	28.126	0.75493	5.46	0.56539	55.254	0.20638	61.305	0.34412
94% Sat.	16.464	0.69074	25.261	0.69014	4.662	0.63089	52.911	0.16904	60.431	0.39654
95% Sat.	15.527	0.65123	23.944	0.70972	4.497	0.6649	51.651	0.21619	59.92	0.51573
96% Sat.	12.803	0.60449	22.049	0.65306	4.467	0.63729	48.305	0.27783	58.541	0.49757
97% Sat.	11.473	0.60236	20.293	0.6573	4.412	0.66283	44.744	0.28685	57.262	0.60007
98% Sat.	9.126	0.57246	19.289	0.66951	4.257	0.65294	39.611	0.45499	54.659	0.85585
99% Sat.	7.927	0.45977	18.773	0.63368	4.24	0.65632	35.806	0.43735	53.205	1.02643
100% Sat.	7.741	0.45866	18.36	0.6451	4.234	0.65638	27.079	0.55288	46.708	1.43304

Table G.9 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 2.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.556	0.24042	1.41	0.30758	1.009	0.33696	0.897	0.29896
92% Sat.	0.44	0.22895	0.886	0.24438	1.003	0.33654	0.735	0.24626
93% Sat.	0.401	0.22672	0.63	0.20287	0.974	0.3371	0.558	0.1705
94% Sat.	0.382	0.23141	0.494	0.21235	0.929	0.33735	0.493	0.15635
95% Sat.	0.381	0.23135	0.451	0.21354	0.929	0.33735	0.464	0.14704
96% Sat.	0.362	0.22869	0.414	0.22385	0.929	0.33735	0.455	0.14368
97% Sat.	0.361	0.22789	0.339	0.19514	0.929	0.33735	0.454	0.1433
98% Sat.	0.327	0.22046	0.309	0.1981	0.904	0.33921	0.436	0.13909
99% Sat.	0.327	0.22046	0.304	0.19976	0.904	0.33921	0.432	0.1376
100% Sat.	0.327	0.22046	0.303	0.20014	0.904	0.33921	0.432	0.13747

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	6.269	0.61874	23.32	1.0847	7.273	0.77419	0.803	0.17353	11.591	0.9697
92% Sat.	4.829	0.59121	21.67	1.10574	4.89	0.63593	0.681	0.17523	9.845	1.00733
93% Sat.	3.585	0.48082	19.997	1.04912	3.517	0.45651	0.525	0.17082	8.174	1.00014
94% Sat.	2.64	0.40306	18.779	1.08969	2.104	0.2444	0.418	0.15939	7.148	0.99346
95% Sat.	2.175	0.38567	18.118	1.12392	1.39	0.18938	0.387	0.16248	6.617	0.98155
96% Sat.	1.72	0.44206	16.774	1.09961	0.971	0.19469	0.37	0.16083	5.642	0.90233
97% Sat.	1.32	0.3453	15.47	1.12772	0.596	0.12807	0.348	0.16117	4.868	0.8268
98% Sat.	1.037	0.33544	14.601	1.11195	0.383	0.04918	0.32	0.16126	4.029	0.73065
99% Sat.	0.939	0.34367	13.7	1.15979	0.295	0.05808	0.316	0.1613	3.523	0.7051
100% Sat.	0.895	0.33202	13.077	1.19281	0.25	0.06545	0.314	0.16131	3.402	0.66752

Table G.10 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 2.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.2387	5.093	0.61208	1.662	0.28086	2.212	0.35454	0.2387
92% Sat.	0.22695	3.624	0.53031	1.624	0.26766	1.833	0.31103	0.22695
93% Sat.	0.23277	2.613	0.42181	1.568	0.25666	1.309	0.29033	0.23277
94% Sat.	0.21274	1.961	0.34875	1.28	0.27035	1.033	0.23674	0.21274
95% Sat.	0.21422	1.635	0.3074	1.28	0.27035	0.957	0.21687	0.21422
96% Sat.	0.22784	1.294	0.2481	1.28	0.27035	0.835	0.23804	0.22784
97% Sat.	0.22643	1.016	0.23156	1.28	0.27035	0.822	0.23228	0.22643
98% Sat.	0.21195	0.885	0.22455	1.222	0.25687	0.747	0.26116	0.21195
99% Sat.	0.21224	0.797	0.21288	1.222	0.25687	0.706	0.26196	0.21224
100% Sat.	0.21224	0.787	0.21078	1.222	0.25687	0.704	0.26051	0.21224

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	20.903	1.11313	30.203	1.40503	32.691	0.95276	23.763	1.55787	41.106	1.3116
92% Sat.	17.468	1.07813	26.708	1.28843	31.016	0.96223	19.746	1.40592	38.009	1.3226
93% Sat.	13.024	1.03624	20.926	1.19595	27.741	1.08442	14.77	1.25842	33.961	1.40167
94% Sat.	9.693	0.95956	16.743	0.8394	25.191	1.09517	10.789	1.00312	30.147	1.3536
95% Sat.	8.049	0.97314	13.544	0.82972	23.844	1.12318	8.382	0.95094	27.141	1.37448
96% Sat.	5.678	0.87275	9.483	0.59203	20.983	1.14207	5.539	0.74004	23.632	1.22533
97% Sat.	4.258	0.71171	6.783	0.56043	18.893	1.16691	3.87	0.55743	21.161	1.1769
98% Sat.	3.303	0.63891	4.734	0.63159	16.796	1.14011	2.652	0.49719	19.324	1.12462
99% Sat.	2.496	0.53553	3.057	0.62994	15.677	1.11667	1.894	0.40704	17.85	1.09176
100% Sat.	2.106	0.48389	2.14	0.57887	15.046	1.07157	1.481	0.37468	17.058	1.08808

Table G.11 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 3.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	4.874	0.38397	9.623	0.86692	3.118	0.27124	5.545	0.76416
92% Sat.	4.526	0.34688	7.664	0.81995	3.073	0.26687	4.905	0.7985
93% Sat.	4.445	0.33534	6.297	0.73131	3.029	0.25447	4.279	0.78989
94% Sat.	3.659	0.31142	4.809	0.66539	2.68	0.23104	3.535	0.63802
95% Sat.	3.623	0.31338	4.342	0.63012	2.68	0.23104	3.327	0.63975
96% Sat.	3.318	0.2722	3.701	0.53424	2.68	0.23104	3.017	0.61678
97% Sat.	3.308	0.27	3.177	0.46887	2.68	0.23104	2.898	0.60524
98% Sat.	3.179	0.25929	2.826	0.46223	2.642	0.22038	2.673	0.59096
99% Sat.	3.176	0.25864	2.682	0.45655	2.642	0.22038	2.527	0.58288
100% Sat.	3.176	0.25868	2.646	0.45827	2.642	0.22038	2.476	0.5929

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	21.048	1.6104	45.782	1.4203	36.824	1.9045	1.241	0.39768	45.573	1.56416
92% Sat.	19.023	1.67132	43.741	1.40524	33.728	1.96371	1.217	0.39293	42.551	1.65222
93% Sat.	14.722	1.94748	40.429	1.66118	28.672	1.62188	1.064	0.40108	38.964	1.59815
94% Sat.	13.71	1.92589	37.46	1.64063	24.091	1.51844	0.924	0.3734	34.857	1.56215
95% Sat.	13.123	1.8674	34.116	1.57654	20.561	1.48552	0.921	0.37314	31.983	1.43175
96% Sat.	11.486	1.82788	30.793	1.39399	15.77	1.18597	0.882	0.37438	28.555	1.34285
97% Sat.	10.785	1.75641	27.689	1.29564	12.439	1.10238	0.831	0.36903	26.012	1.18991
98% Sat.	10.094	1.68147	24.849	1.20169	9.624	0.77985	0.773	0.36786	23.698	1.08293
99% Sat.	9.907	1.66813	22.412	1.1419	7.364	0.66775	0.749	0.36996	21.933	0.94078
100% Sat.	9.885	1.66734	21.034	1.14966	6.135	0.54692	0.734	0.37032	20.999	0.85195

Table G.12 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 3.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	5.67	0.84011	13.247	0.5294	3.864	0.4235	8.055	0.81991
92% Sat.	5.26	0.70803	10.587	0.46555	3.739	0.41644	7.138	0.79176
93% Sat.	4.631	0.67288	8.102	0.39619	3.586	0.42001	6.322	0.73552
94% Sat.	3.857	0.63937	6.097	0.33249	3.31	0.3937	5.048	0.70633
95% Sat.	3.826	0.62633	5.386	0.32252	3.31	0.3937	4.77	0.6464
96% Sat.	3.584	0.66142	4.549	0.30103	3.31	0.3937	4.349	0.54042
97% Sat.	3.481	0.58973	3.553	0.34854	3.31	0.3937	3.737	0.51988
98% Sat.	3.167	0.53058	2.924	0.37268	3.267	0.3934	3.392	0.46844
99% Sat.	3.161	0.52706	2.696	0.38381	3.267	0.3934	3.178	0.44821
100% Sat.	3.159	0.52716	2.63	0.40129	3.267	0.3934	2.886	0.43284

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	41.113	0.87998	8.696	0.69597	15.769	1.21692	2.093	0.66217	44.265	1.39103
92% Sat.	39.203	0.86035	7.333	0.73853	15.164	1.15416	2.009	0.63457	41.233	1.5285
93% Sat.	36.783	0.81713	5.327	0.64357	12.167	1.19016	1.862	0.62443	35.673	1.55053
94% Sat.	34.413	0.70478	3.974	0.61924	10.502	1.17688	1.677	0.62775	30.338	1.4566
95% Sat.	33.078	0.68227	3.693	0.62739	10.439	1.13135	1.669	0.62624	26.82	1.4103
96% Sat.	31.35	0.64377	3.116	0.64582	8.017	1.06955	1.609	0.62768	21.453	1.24042
97% Sat.	29.787	0.6584	2.661	0.63965	6.914	0.87445	1.47	0.62475	17.227	1.1325
98% Sat.	28.05	0.59985	2.358	0.66677	6.137	0.7701	1.397	0.62153	13.333	0.94834
99% Sat.	25.772	0.52418	2.26	0.6697	5.839	0.74021	1.334	0.60836	10.573	0.90927
100% Sat.	24.428	0.50737	2.138	0.65433	5.7	0.63742	1.236	0.59493	8.673	0.81527

Table G.13 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 4.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	7.41	0.91867	14.981	1.40107	3.45	0.59756	9.407	0.9532
92% Sat.	6.89	0.8832	12.035	1.47644	3.344	0.58923	8.334	0.83759
93% Sat.	6.166	0.84322	9.258	1.20551	3.228	0.5558	7.278	0.87614
94% Sat.	5.09	0.76615	6.695	1.19748	3.009	0.55938	5.824	0.83686
95% Sat.	5.013	0.7614	5.899	1.10566	3.009	0.55938	5.441	0.72425
96% Sat.	4.647	0.66867	4.651	0.93847	3.009	0.55938	4.748	0.70319
97% Sat.	4.49	0.67177	3.691	0.76293	3.009	0.55938	4.11	0.74643
98% Sat.	4.107	0.66613	2.93	0.6899	2.979	0.56125	3.598	0.73364
99% Sat.	4.097	0.6627	2.62	0.64934	2.979	0.56125	3.278	0.69707
100% Sat.	4.093	0.66222	2.585	0.64098	2.979	0.56125	2.974	0.73911

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	30.461	1.20715	5.131	0.43549	41.689	0.62006	1.891	0.30108	52.439	1.24597
92% Sat.	29.318	1.17879	4.231	0.47158	39.912	0.61454	1.718	0.34746	51.214	1.22141
93% Sat.	28.059	1.05483	3.159	0.46351	37.623	0.56508	1.519	0.34553	49.33	1.12174
94% Sat.	26.402	1.03324	2.512	0.45984	35.671	0.5669	1.417	0.33492	46.658	1.06431
95% Sat.	25.527	0.95952	2.291	0.4471	34.288	0.58379	1.415	0.33469	44.03	1.03637
96% Sat.	23.939	0.83564	1.946	0.44834	32.58	0.5242	1.392	0.33533	40.795	0.75373
97% Sat.	22.506	0.7828	1.642	0.43737	30.536	0.43593	1.354	0.3496	37.933	0.74957
98% Sat.	21.828	0.72001	1.476	0.44318	28.315	0.48718	1.309	0.34445	34.526	0.47286
99% Sat.	20.361	0.64859	1.383	0.43044	25.723	0.46615	1.307	0.344	31.742	0.49138
100% Sat.	19.287	0.60273	1.247	0.4373	24.528	0.42652	1.307	0.344	29.42	0.48352

Table G.14 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 4.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	9.975	0.82528	19.397	1.28929	4.23	0.21372	12.215	0.80294
92% Sat.	9.203	0.8409	16.037	1.36004	4.132	0.19687	10.939	0.76933
93% Sat.	8.053	0.91047	12.419	1.13794	4.031	0.19611	9.75	0.74933
94% Sat.	7.179	0.77078	9.683	0.85626	3.782	0.18683	8.086	0.81531
95% Sat.	7.024	0.73317	8.538	0.86487	3.782	0.18683	7.584	0.82613
96% Sat.	6.403	0.71108	6.415	0.75969	3.782	0.18683	6.743	0.79367
97% Sat.	5.955	0.62474	5.106	0.57305	3.782	0.18683	5.711	0.77177
98% Sat.	5.385	0.52646	3.873	0.48607	3.719	0.18147	4.993	0.68088
99% Sat.	5.353	0.51825	3.407	0.44412	3.719	0.18147	4.487	0.64293
100% Sat.	5.352	0.5182	3.34	0.41981	3.719	0.18147	3.78	0.52247

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	41.4	0.88031	48.754	0.81776	36.607	0.6465	2.757	0.45405	55.817	0.82299
92% Sat.	38.764	0.91478	47.296	0.78635	35.613	0.61165	2.567	0.4416	54.28	0.90603
93% Sat.	34.198	0.88204	43.521	0.87833	33.246	0.52702	2.315	0.48824	51.906	1.04288
94% Sat.	30.038	0.72726	39.489	0.86717	31.404	0.47994	2.074	0.46858	48.953	0.87782
95% Sat.	27.911	0.6538	36.52	0.81792	29.889	0.4822	2.041	0.47354	46.179	0.86393
96% Sat.	24.369	0.63002	31.493	0.70596	26.824	0.51988	1.956	0.45768	42.475	0.90346
97% Sat.	21.31	0.56337	27.252	0.85795	23.862	0.62421	1.751	0.48621	39.028	0.88595
98% Sat.	18.842	0.56527	22.11	0.82516	20.949	0.59265	1.64	0.47778	34.3	0.73326
99% Sat.	16.195	0.53462	18.426	0.80079	18.311	0.69433	1.565	0.46499	30.977	0.7239
100% Sat.	14.731	0.49277	15.229	0.79795	17.305	0.74217	1.337	0.4747	27.815	0.62182

Table G.15 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 5.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	10.94	1.2603	23.364	0.56352	4.983	0.45175	13.884	0.73928
92% Sat.	10.203	1.29177	19.853	0.56297	4.848	0.45968	12.492	0.73813
93% Sat.	8.96	1.22261	15.944	0.57053	4.715	0.44832	11.325	0.76415
94% Sat.	8.157	1.10044	12.868	0.4678	4.533	0.49744	9.571	0.6887
95% Sat.	7.949	1.08448	11.581	0.52667	4.533	0.49744	8.9	0.62392
96% Sat.	7.373	0.99215	8.858	0.51712	4.533	0.49744	7.865	0.61737
97% Sat.	6.891	0.91897	7.391	0.52725	4.533	0.49744	6.779	0.51103
98% Sat.	6.337	0.79852	5.613	0.54724	4.473	0.50432	5.912	0.47017
99% Sat.	6.233	0.75932	4.888	0.61904	4.473	0.50432	5.266	0.45717
100% Sat.	6.23	0.76012	4.808	0.62746	4.473	0.50432	4.281	0.36841

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	29.117	0.78236	40.891	0.28662	52.338	1.08782	3.765	0.48145	37.247	0.46983
92% Sat.	27.097	0.80326	39.925	0.28697	50.895	1.1057	3.406	0.46308	36.192	0.43563
93% Sat.	23.239	0.95542	38.845	0.3219	47.121	1.00008	2.913	0.42407	34.642	0.42843
94% Sat.	19.94	0.87342	37.283	0.36163	43.403	0.94068	2.499	0.38754	33.019	0.50808
95% Sat.	18.487	0.86523	36.302	0.35525	40.687	0.99341	2.388	0.39764	32.04	0.48048
96% Sat.	15.406	0.88716	34.723	0.34128	35.594	0.87966	2.257	0.40014	30.368	0.48104
97% Sat.	13.029	0.826	33.091	0.35566	31.388	0.83024	2.08	0.40569	28.808	0.51028
98% Sat.	11.082	0.73368	31.818	0.44184	25.768	0.70829	1.943	0.40333	27.398	0.52576
99% Sat.	9.31	0.68304	29.342	0.4764	21.862	0.69288	1.862	0.39054	25.21	0.47825
100% Sat.	8.491	0.65042	28.349	0.45456	17.984	0.65929	1.799	0.39588	24.36	0.40894

Table G.16 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 5.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	10.841	1.19422	24.656	0.96927	5.123	0.38787	14.836	0.78122
92% Sat.	9.585	1.11272	20.66	1.20173	5.079	0.37831	13.337	0.79383
93% Sat.	8.642	1.04346	16.829	1.08635	5.016	0.37122	12.04	0.82875
94% Sat.	7.569	1.01516	13.518	1.14632	4.906	0.37449	9.99	0.89309
95% Sat.	7.434	1.01282	12.221	1.06888	4.906	0.37449	9.262	0.84111
96% Sat.	7.011	0.90844	9.28	1.01732	4.906	0.37449	8.321	0.81732
97% Sat.	6.764	0.88889	7.821	0.8969	4.906	0.37449	6.932	0.68129
98% Sat.	6.172	0.76832	5.751	0.95285	4.843	0.3717	6.064	0.61198
99% Sat.	6.141	0.77269	4.905	0.93524	4.843	0.3717	5.552	0.54203
100% Sat.	6.138	0.77199	4.807	0.93424	4.843	0.3717	4.422	0.40546

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	52.432	0.89481	44.198	0.74026	51.306	0.86014	52.352	0.79188	57.204	0.7569
92% Sat.	51.305	0.95166	43.149	0.71233	50.18	0.88049	50.783	0.87654	55.894	0.81016
93% Sat.	48.068	1.00548	40.818	0.79626	47.114	1.04852	46.799	1.0891	53.01	0.73601
94% Sat.	44.852	0.96322	38.127	0.81853	43.969	1.08001	42.997	1.1891	49.824	0.76957
95% Sat.	42.958	1.01925	35.721	0.86751	41.524	1.12917	40.349	1.30283	47.406	0.86729
96% Sat.	38.693	1.06021	32.09	1.03135	37.006	1.28125	35.296	1.33009	42.887	1.13876
97% Sat.	34.963	0.94985	28.518	1.00943	33.037	1.17475	30.719	1.17239	38.177	0.96604
98% Sat.	29.832	0.95748	23.647	1.10441	27.661	1.17624	25.227	0.95457	32.315	1.02431
99% Sat.	27.157	0.95417	19.702	0.9256	23.67	1.06675	21.249	0.75225	28.075	0.87948
100% Sat.	23.159	0.83433	15.357	0.73654	19.24	0.85291	17.153	0.44869	22.844	0.76806

Table G.17 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 2.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	0.56	0.20161	0.28	0.17663	0.357	0.22678	3.352	0.65693
92% Sat.	0.541	0.20071	0.236	0.17498	0.308	0.20104	2.55	0.56664
93% Sat.	0.464	0.19726	0.185	0.16737	0.231	0.16643	1.832	0.42951
94% Sat.	0.464	0.19726	0.181	0.16819	0.231	0.16643	1.662	0.39373
95% Sat.	0.464	0.19726	0.171	0.16842	0.231	0.16643	1.384	0.36281
96% Sat.	0.464	0.19726	0.161	0.17104	0.231	0.16643	1.046	0.31313
97% Sat.	0.464	0.19726	0.159	0.17144	0.231	0.16643	0.908	0.32619
98% Sat.	0.464	0.19726	0.159	0.17169	0.231	0.16643	0.832	0.27754
99% Sat.	0.464	0.19726	0.159	0.17141	0.231	0.16643	0.788	0.27967
100% Sat.	0.464	0.19726	0.159	0.17141	0.231	0.16643	0.787	0.27965

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	5.487	0.8077	5.263	0.2602	2.117	0.29741	2.17	0.44116	2.774	0.52405
92% Sat.	4.497	0.74411	4.881	0.27769	1.805	0.30752	1.649	0.33169	2.467	0.55679
93% Sat.	3.704	0.65787	4.208	0.23328	1.323	0.22912	1.304	0.25215	1.989	0.46615
94% Sat.	3.335	0.64918	3.993	0.23801	1.213	0.22375	1.176	0.27472	1.888	0.45537
95% Sat.	2.682	0.57211	3.609	0.23879	0.94	0.21074	1.094	0.2743	1.594	0.44152
96% Sat.	2.029	0.47271	2.767	0.24889	0.708	0.2021	0.961	0.23374	1.349	0.42464
97% Sat.	1.864	0.49041	2.533	0.28267	0.658	0.21175	0.938	0.2364	1.3	0.43549
98% Sat.	1.728	0.46429	2.423	0.26752	0.567	0.19479	0.928	0.2341	1.209	0.40535
99% Sat.	1.639	0.45218	2.392	0.26532	0.543	0.19441	0.925	0.23631	1.184	0.40539
100% Sat.	1.629	0.44942	2.392	0.26515	0.542	0.19447	0.921	0.23815	1.184	0.40525

Table G.18 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 2.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	1.249	0.35869	0.985	0.18094	0.654	0.1806	9.115	0.70305
92% Sat.	1.162	0.35048	0.799	0.17284	0.55	0.18053	7.409	0.63936
93% Sat.	0.712	0.31938	0.413	0.14454	0.29	0.16047	5.745	0.67303
94% Sat.	0.712	0.31938	0.323	0.11978	0.29	0.16047	5.031	0.60196
95% Sat.	0.712	0.31938	0.273	0.11278	0.29	0.16047	4.06	0.5092
96% Sat.	0.712	0.31971	0.211	0.10204	0.29	0.16047	2.8	0.4122
97% Sat.	0.712	0.31971	0.196	0.09757	0.29	0.16047	2.173	0.31543
98% Sat.	0.712	0.31971	0.189	0.0965	0.29	0.16047	1.924	0.32533
99% Sat.	0.712	0.31971	0.187	0.0952	0.29	0.16047	1.651	0.31497
100% Sat.	0.712	0.31971	0.186	0.09523	0.29	0.16047	1.576	0.31518

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	9.878	1.43419	16.56	0.50846	7.379	0.9081	7.115	0.56543	15.613	1.07808
92% Sat.	9.216	1.43918	16.265	0.52103	6.614	0.87152	6.579	0.53516	14.885	1.10767
93% Sat.	7.757	1.40239	15.526	0.60103	4.926	0.76336	4.936	0.50689	13.371	1.07603
94% Sat.	7.206	1.32484	15.415	0.59817	4.392	0.7084	4.567	0.51574	13.048	1.08145
95% Sat.	6.576	1.16777	15.289	0.6002	3.763	0.5026	4.306	0.40273	12.413	0.97245
96% Sat.	5.181	1.08574	14.275	0.55528	2.424	0.43705	2.545	0.34026	11	0.9787
97% Sat.	4.641	0.97694	13.867	0.54436	1.955	0.38857	1.826	0.27631	10.399	0.90659
98% Sat.	4.204	0.87145	13.64	0.49912	1.557	0.31487	1.446	0.27467	9.915	0.86072
99% Sat.	4.062	0.78305	13.627	0.49431	1.432	0.2481	1.417	0.26365	9.781	0.79376
100% Sat.	4.06	0.78291	13.627	0.49431	1.43	0.24782	1.416	0.26338	9.78	0.79409

Table G.19 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 3.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	3.921	0.49988	3.445	0.26138	2.202	0.37658	18.99	1.21448
92% Sat.	3.833	0.46402	3.041	0.24685	2.107	0.37415	16.705	1.2513
93% Sat.	3.379	0.3761	2.423	0.27247	1.848	0.34642	14.431	1.17617
94% Sat.	3.375	0.37695	2.176	0.31993	1.848	0.34642	12.955	1.09688
95% Sat.	3.358	0.3505	1.911	0.32969	1.848	0.34642	10.959	0.92249
96% Sat.	3.347	0.35494	1.568	0.32497	1.848	0.34642	8.836	0.73313
97% Sat.	3.347	0.35486	1.497	0.32209	1.848	0.34642	7.605	0.74253
98% Sat.	3.347	0.35486	1.47	0.32345	1.848	0.34642	7.018	0.61141
99% Sat.	3.347	0.35486	1.451	0.32761	1.848	0.34642	6.285	0.55009
100% Sat.	3.347	0.35486	1.451	0.32753	1.848	0.34642	6.057	0.53284

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	18.498	1.02662	14.591	0.62991	22.008	0.69901	16.742	1.08411	23.357	0.61542
92% Sat.	17.481	0.9534	14.07	0.6038	21.339	0.65282	15.833	1.04895	22.67	0.6024
93% Sat.	15.458	0.88942	12.381	0.72225	20.168	0.58855	13.829	1.00072	21.19	0.52322
94% Sat.	14.484	0.91	11.717	0.711	19.412	0.63587	12.884	0.98056	20.635	0.51202
95% Sat.	13.163	0.73359	11.124	0.6409	18.473	0.7056	11.594	0.81153	19.904	0.46594
96% Sat.	10.678	0.7203	8.537	0.6533	15.969	0.85661	9.085	0.75092	18.017	0.46319
97% Sat.	9.48	0.68787	7.314	0.5706	14.623	0.85881	7.886	0.686	17.026	0.50859
98% Sat.	8.513	0.71836	6.705	0.6238	13.649	0.85431	6.903	0.72623	16.15	0.49176
99% Sat.	8.107	0.62506	6.673	0.61769	13.44	0.82578	6.535	0.61085	15.969	0.51619
100% Sat.	8.103	0.6256	6.669	0.61766	13.438	0.82664	6.532	0.61138	15.965	0.51538

Table G.20 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 3.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	4.035	0.37867	5.876	0.73465	2.677	0.29648	24.727	1.8319
92% Sat.	3.914	0.38389	5.106	0.74637	2.597	0.28401	22.391	1.90861
93% Sat.	3.265	0.38018	3.582	0.68844	2.249	0.2823	19.826	1.94652
94% Sat.	3.258	0.38508	3.168	0.6648	2.249	0.2823	17.878	1.9307
95% Sat.	3.217	0.37876	2.563	0.64594	2.249	0.2823	15.319	1.72412
96% Sat.	3.198	0.37849	1.939	0.55494	2.249	0.2823	12.607	1.80366
97% Sat.	3.196	0.37626	1.699	0.54023	2.249	0.2823	10.404	1.67591
98% Sat.	3.196	0.37626	1.6	0.52847	2.249	0.2823	9.246	1.66819
99% Sat.	3.196	0.37626	1.53	0.52402	2.249	0.2823	8.077	1.4302
100% Sat.	3.196	0.37626	1.528	0.52471	2.249	0.2823	7.1	1.35039

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	16.833	0.64696	23.853	0.9356	16.906	0.59382	8.538	0.73826	26.973	0.62733
92% Sat.	16.117	0.64048	23.379	0.90102	16.289	0.56686	7.369	0.73251	25.94	0.62925
93% Sat.	14.205	0.75539	21.964	0.80111	14.259	0.5423	6.273	0.65678	24.12	0.6352
94% Sat.	13.95	0.72001	21.526	0.82083	13.582	0.52514	5.206	0.57982	23.42	0.54402
95% Sat.	12.994	0.719	21.156	0.72869	12.826	0.48599	4.445	0.5493	22.179	0.52761
96% Sat.	11.157	0.64749	19.537	0.71784	10.215	0.51081	3.434	0.53637	20.271	0.5021
97% Sat.	10.098	0.5872	18.625	0.70637	8.751	0.46685	3.098	0.55797	19.35	0.5313
98% Sat.	9.563	0.50597	17.922	0.6194	7.902	0.45307	2.384	0.44679	18.454	0.48761
99% Sat.	9.402	0.50179	17.841	0.61029	7.75	0.48126	2.341	0.45335	18.02	0.46823
100% Sat.	9.4	0.49979	17.831	0.60425	7.741	0.4838	1.97	0.41527	18.01	0.46651

Table G.21 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 4.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	3.292	0.30295	7.483	0.79063	2.732	0.30581	29.629	2.07513
92% Sat.	3.207	0.28771	6.54	0.76191	2.636	0.30979	27.405	2.10963
93% Sat.	2.811	0.32803	4.793	0.67741	2.252	0.31136	24.444	1.98061
94% Sat.	2.78	0.31016	4.103	0.61234	2.252	0.31136	22.252	1.86258
95% Sat.	2.749	0.29441	3.126	0.53047	2.252	0.31136	19.084	1.80597
96% Sat.	2.73	0.29265	2.283	0.47636	2.252	0.31136	15.981	1.71031
97% Sat.	2.711	0.2837	1.731	0.43411	2.252	0.31136	13.123	1.56663
98% Sat.	2.711	0.28369	1.489	0.42228	2.252	0.31136	11.238	1.27009
99% Sat.	2.711	0.28369	1.308	0.40752	2.252	0.31136	9.808	1.05408
100% Sat.	2.711	0.28369	1.298	0.40549	2.252	0.31136	8.006	0.90451

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	29.442	0.83417	22.368	1.39738	16.294	1.11831	11.998	1.26426	25.519	1.16973
92% Sat.	28.413	0.86571	19.736	1.33836	16.023	1.132	10.312	1.18457	24.37	1.21446
93% Sat.	26.61	0.84476	16.96	1.29103	14.656	0.99586	8.85	1.11068	21.913	1.20591
94% Sat.	25.935	0.88401	14.751	1.28332	14.075	0.9824	7.462	1.10425	20.719	1.16274
95% Sat.	24.454	0.81652	12.231	1.13973	13.554	0.95362	6.247	1.00471	18.727	1.13237
96% Sat.	22.93	0.78824	9.543	0.99381	10.534	0.88055	4.954	0.97719	15.428	1.07817
97% Sat.	21.931	0.77551	7.809	0.8823	9.523	0.82296	4.288	0.88043	13.429	0.97325
98% Sat.	20.977	0.74079	6.189	0.72439	8.884	0.50326	3.123	0.69707	11.867	0.91519
99% Sat.	20.552	0.6082	5.737	0.68019	8.883	0.50151	3.016	0.69334	11.047	0.72672
100% Sat.	20.542	0.6071	4.711	0.60469	8.883	0.50151	2.207	0.50189	11.008	0.72021

Table G.22 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 4.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	3.273	0.48132	10.134	0.84094	2.885	0.32612	33.379	2.75487
92% Sat.	3.148	0.45001	9.09	0.83711	2.829	0.33028	31.494	2.68547
93% Sat.	2.772	0.40988	6.716	0.71941	2.564	0.36107	28.835	2.75324
94% Sat.	2.721	0.38791	5.943	0.65046	2.564	0.36107	26.636	2.58034
95% Sat.	2.693	0.38777	4.47	0.63489	2.564	0.36107	23.487	2.60693
96% Sat.	2.677	0.38966	2.943	0.57601	2.564	0.36107	19.987	2.47635
97% Sat.	2.675	0.38985	2.033	0.55798	2.564	0.36107	16.945	2.46137
98% Sat.	2.675	0.38985	1.638	0.50695	2.564	0.36107	14.552	2.25869
99% Sat.	2.675	0.38985	1.287	0.52188	2.564	0.36107	12.634	2.12094
100% Sat.	2.675	0.38985	1.265	0.52256	2.564	0.36107	10.196	2.11471

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	17.52	0.95958	22.477	1.38951	28.39	0.42692	2.671	0.6073	28.518	0.6397
92% Sat.	17.332	0.90985	20.235	1.27799	27.617	0.39098	2.385	0.6153	27.973	0.62638
93% Sat.	16.539	0.8467	17.639	1.38811	26.144	0.37081	1.856	0.62867	26.891	0.61542
94% Sat.	16.501	0.84685	15.874	1.34871	25.541	0.37462	1.785	0.62612	26.455	0.6333
95% Sat.	16.154	0.86821	13.359	1.23919	24.274	0.37553	1.752	0.61759	25.516	0.6317
96% Sat.	15.408	0.83772	10.942	1.15282	22.497	0.44907	1.627	0.57885	23.759	0.73031
97% Sat.	15.117	0.75398	9.115	1.1355	21.376	0.52432	1.422	0.54387	22.859	0.66696
98% Sat.	14.943	0.65974	7.272	1.02201	20.218	0.61649	1.378	0.52752	21.934	0.65819
99% Sat.	14.921	0.67066	6.753	0.9962	19.743	0.6125	1.375	0.52702	21.709	0.59635
100% Sat.	14.92	0.67046	5.116	0.94314	19.735	0.61265	1.375	0.52702	21.69	0.59383

Table G.23 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 5.0)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	4.061	0.66353	14.376	1.24691	3.54	0.17036	37.622	1.17569
92% Sat.	3.969	0.65365	13.257	1.217	3.503	0.17352	35.911	1.36498
93% Sat.	3.658	0.6253	10.638	1.19952	3.235	0.18384	33.377	1.40182
94% Sat.	3.577	0.62424	9.634	1.14346	3.235	0.18384	31.154	1.47245
95% Sat.	3.575	0.62436	7.809	1.11666	3.235	0.18384	28.048	1.45345
96% Sat.	3.547	0.61672	5.313	0.99381	3.235	0.18384	24.445	1.45624
97% Sat.	3.546	0.61711	3.986	0.96968	3.235	0.18384	21.283	1.42938
98% Sat.	3.546	0.61711	3.253	1.00997	3.235	0.18384	18.446	1.24814
99% Sat.	3.546	0.61711	2.633	0.98732	3.235	0.18384	16.359	1.14356
100% Sat.	3.546	0.61711	2.605	0.99042	3.235	0.18384	13.29	1.13912

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	42.083	1.77859	27.556	0.46988	30.087	0.4652	25.643	1.10566	33.169	0.81695
92% Sat.	40.998	1.65848	26.937	0.49599	29.255	0.49281	23.677	1.10556	32.242	0.83352
93% Sat.	39.108	1.74223	25.429	0.59371	27.609	0.56401	21.195	1.07123	30.552	0.98144
94% Sat.	37.585	1.73734	24.56	0.65921	27.221	0.56762	19.263	1.0327	29.92	0.99653
95% Sat.	34.78	1.5842	23.062	0.73163	25.83	0.50733	16.612	1.01147	28.492	1.15982
96% Sat.	31.179	1.69079	19.964	0.81781	23.885	0.54813	13.901	0.94708	26.038	1.19548
97% Sat.	28.275	1.51116	18.07	0.8382	22.535	0.52678	11.973	0.86109	24.459	1.23009
98% Sat.	25.366	1.47826	16.183	0.74686	21.456	0.50621	9.715	0.74826	22.719	1.17753
99% Sat.	22.497	1.32026	15.228	0.61316	20.618	0.43384	8.841	0.75994	21.788	1.06824
100% Sat.	20.658	1.16297	15.183	0.61234	20.602	0.43431	6.303	0.6134	21.742	1.06545

Table G.24 Percentages of TTIs (Mean and STD) in the exploitation stage for the GBR user satisfaction levels based on NAUT-MMF with the static windowing factor: $(\rho = 5.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Policies / GBR Satisfaction	GPF-BF [Mean]	GPF-BF [STD]	GPF-RAD [Mean]	GPF-RAD [STD]	GPF-Min/Max [Mean]	GPF-Min/Max [STD]	GPF-LM [Mean]	GPF-LM [STD]
91% Sat.	4.1	0.72895	15.853	1.30222	3.688	0.37266	38.984	1.08913
92% Sat.	3.913	0.66355	14.731	1.27608	3.672	0.37442	37.52	1.03052
93% Sat.	3.563	0.68665	11.933	1.32536	3.506	0.43096	35.073	1.14007
94% Sat.	3.413	0.61848	10.716	1.27253	3.506	0.43096	32.887	1.09712
95% Sat.	3.412	0.61917	8.757	1.22012	3.506	0.43096	30.146	1.12868
96% Sat.	3.363	0.57656	6.068	1.12264	3.506	0.43096	26.421	1.08437
97% Sat.	3.347	0.56432	4.723	1.09383	3.506	0.43096	23.229	1.21325
98% Sat.	3.344	0.56473	3.72	1.02721	3.506	0.43096	20.017	1.06148
99% Sat.	3.344	0.56473	2.989	0.96041	3.506	0.43096	17.758	0.93309
100% Sat.	3.344	0.56473	2.947	0.96632	3.506	0.43096	13.801	1.1775

Policies / GBR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	29.607	0.47136	21.205	0.52117	26.745	0.99674	31.803	0.60002	34.354	0.85734
92% Sat.	28.907	0.42047	20.6	0.49336	26.149	0.97219	30.915	0.58234	33.544	0.8327
93% Sat.	27.293	0.46821	19.629	0.48021	24.878	0.93112	29.015	0.53047	31.744	0.85698
94% Sat.	26.395	0.56007	19.084	0.49801	24.191	0.9324	28.02	0.59943	30.826	0.93346
95% Sat.	24.537	0.5214	17.569	0.58742	22.617	0.84569	25.734	0.63108	28.544	0.95327
96% Sat.	21.172	0.59211	14.949	0.62598	19.382	0.88745	22.043	0.63572	25.032	1.10549
97% Sat.	19.083	0.59616	13.222	0.65426	17.326	0.88707	19.78	0.64489	22.819	1.14091
98% Sat.	17.088	0.6081	11.306	0.65028	15.313	0.9125	17.647	0.62578	20.59	1.06487
99% Sat.	16.18	0.58475	10.322	0.63099	14.302	0.57282	15.921	0.52404	18.765	0.87393
100% Sat.	16.122	0.58338	10.277	0.62584	14.26	0.57134	15.805	0.5282	18.629	0.87464

G.3 Percentages of TTIs for the GBR Testing Rewards Based on Infinite Buffer, CBR and VBR Traffic Types

The scheduler reward model is presented in Sub-section 6.3.3, Chapter 6 and represents the sum of sub-rewards received from each active user. It is important to set the reward function as a difference between two consecutive rewards (Eq. 6.30) in order to highlight the advantage of using certain rule between two consecutive TTIs. In this sense, the parameter that sets this characteristic is $\ell_G = 1$. The role of the testing rewards is very important since the windowing factor optimality can be decided based on the mean percentage of TTIs when the rewards are moderate, punishment or maximized. When the mean percentage of TTIs with maximum rewards $\overline{p}_{TTI}^{G,MRW}$ is relatively high when compared with other percentages of $\overline{p}_{TTI}^{G,PSH}$ and $\overline{p}_{TTI}^{G,mRW}$, then the optimum range of windowing factor is reached. Otherwise, the scheduling policies remain sub-optimal and the mean percentage $\overline{p}_{TTI}^{G,MRW}$ becomes very small when compared with $\overline{p}_{TTI}^{G,PSH}$ or $\overline{p}_{TTI}^{G,mRW}$. The rest of this section is organized as follows: Tables G.25 to G.32 list the mean percentage of TTIs for the infinite buffer traffic type, Tables G.33 to G.40 present the performance of the testing rewards of different scheduling policies under the CBR traffic type and finally, in Tables G. 41 to G.48 the impact of the VBR traffic type is studied in terms of the mean percentage of TTIs with different reward types. Each table presents the best and worst options of the scheduling policies for the mean percentage of TTIs when the rewards are moderate $\left(\mathcal{RW}_t^G > 0\right)$, punishment $\left(\mathcal{RW}_t^G < 0\right)$ or maximized $\left(\mathcal{RW}_t^G = 1\right)$. The sustainable scheduling policies for the optimum windowing factors involved in the AUT-MMF computations are obtained if the mean percentages of TTIs for the entire domain of GBR levels are maximized when compared with other existing techniques. At the same time, the mean percentages of TTIs with moderate and punishment rewards must be as small as possible. Additionally, the STD values must be minimized in order to prove the sustainability of the obtained policies.

Table G.25 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 2.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	41.075	0.13464	57.258	0.48515	1.667	0.48855
QV2	39.905	1.06773	60.053	1.07322	0.042	0.03055
QVMAX	39.234	1.04901	60.742	1.05302	0.024	0.01582
QVMAX2	41.058	0.14076	57.331	0.43968	1.611	0.4829
ACLA	41.737	0.18432	58.246	0.18592	0.017	0.007
Best	39.234	**QVMAX**	57.258	**QV**	1.667	**QV**
Worst	41.737	**ACLA**	60.742	**QVMAX**	0.017	**ACLA**

Table G.26 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 2.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	38.798	0.26223	54.499	0.65604	6.704	0.8012
QV2	39.458	0.68357	60.411	0.7058	0.131	0.06128
QVMAX	37.95	0.24259	54.071	0.63079	7.979	0.77038
QVMAX2	41.285	0.71605	58.031	0.75114	0.684	0.14737
ACLA	37.827	0.39771	54.772	0.55957	7.401	0.73288
Best	37.827	**ACLA**	54.071	**QVMAX**	7.979	**QVMAX**
Worst	41.285	**QVMAX2**	60.411	**QV2**	0.131	**QV2**

Table G.27 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 3.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	39.383	0.19072	55.388	0.5424	5.228	0.63971
QV2	40.559	0.11838	55.597	0.59811	3.844	0.59863
QVMAX	40.245	0.17724	59.662	0.18704	0.093	0.02781
QVMAX2	33.676	0.28772	48.97	1.38969	17.355	1.55073
ACLA	33.546	0.35434	48.355	1.42618	18.1	1.70409
Best	33.546	**ACLA**	48.355	**ACLA**	17.355	**QVMAX2**
Worst	40.559	**QV2**	59.662	**QVMAX**	0.093	**QVMAX**

Table G.28 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 3.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	37.833	0.48343	54.904	1.07035	7.262	1.51712
QV2	39.622	0.16706	60.236	0.17666	0.142	0.02449
QVMAX	39.564	0.24988	60.291	0.25774	0.145	0.02311
QVMAX2	40.656	0.83019	59.081	0.82192	0.264	0.06883
ACLA	41.269	0.18662	56.049	0.41787	2.682	0.45031
Best	37.833	**QV**	54.904	**QV**	7.262	**QV**
Worst	41.269	**ACLA**	60.291	**QVMAX**	0.142	**QV2**

Table G.29 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 4.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	37.426	0.24462	54.056	0.37984	8.519	0.41931
QV2	35.928	0.2004	54.519	0.50922	9.552	0.53912
QVMAX	27.402	0.45724	39.097	1.80471	33.501	2.21634
QVMAX2	39.033	0.15674	53.033	0.57563	7.934	0.5573
ACLA	27.81	0.56035	40.083	1.56634	32.106	2.06566
Best	27.81	**ACLA**	40.083	**ACLA**	33.501	**QVMAX**
Worst	39.033	**QVMAX2**	54.519	**QV2**	7.934	**QVMAX2**

Table G.30 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 4.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	40.573	0.21593	58.516	0.36789	0.911	0.24587
QV2	40.269	0.12539	59.444	0.14737	0.287	0.04427
QVMAX	26.993	0.5481	54.194	3.01434	18.813	3.44769
QVMAX2	40.371	0.27402	59.314	0.2967	0.315	0.05038
ACLA	24.82	0.60499	36.385	1.14607	38.795	1.5968
Best	24.82	**ACLA**	36.385	**ACLA**	38.795	**ACLA**
Worst	40.573	**QV**	59.444	**QV2**	0.287	**QV2**

Table G.31 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 5.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	25.113	0.6183	43.686	1.00904	31.2	1.3189
QV2	37.647	0.25141	58.861	0.51344	3.493	0.32163
QVMAX	40.822	0.25227	58.519	0.37507	0.66	0.19632
QVMAX2	33.266	0.17372	52.717	0.88038	14.016	0.85303
ACLA	34.034	0.58005	50.495	0.42328	15.471	0.93033
Best	25.113	**QV**	43.686	**QV**	31.2	**QV**
Worst	40.822	**QVMAX**	58.861	**QV2**	0.66	**QVMAX**

Table G.32 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 5.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Infinite Buffer

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	37.939	0.17418	54.323	0.52047	7.737	0.45918
QV2	31.744	0.1694	49.918	0.5222	18.338	0.64376
QVMAX	29.01	0.19505	66.757	0.65244	4.233	0.65654
QVMAX2	29.79	0.19434	43.237	0.64874	26.972	0.55484
ACLA	21.628	0.32498	31.799	1.17593	46.573	1.44412
Best	21.628	**ACLA**	31.799	**ACLA**	46.573	**ACLA**
Worst	37.939	**QV**	54.323	**QV**	4.233	**QVMAX**

Table G.33 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 2.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	43.607	0.39025	55.556	0.40623	0.835	0.33253
QV2	43.756	0.52298	43.231	1.13213	13.013	1.19285
QVMAX	41.553	0.56192	58.257	0.55475	0.189	0.06441
QVMAX2	38.754	0.44386	60.68	0.56824	0.564	0.18231
ACLA	42.156	0.56456	54.498	1.03422	3.345	0.66994
Best	38.754	**QVMAX2**	43.231	**QV2**	13.013	**QV2**
Worst	43.756	**QV2**	60.68	**QVMAX2**	0.189	**QVMAX**

Table G.34 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 2.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	41.74	0.33347	56.222	0.58721	2.038	0.48426
QV2	40.407	0.80935	57.508	1.2466	2.084	0.57909
QVMAX	36.038	0.36624	48.98	1.02305	14.981	1.07165
QVMAX2	39.623	0.63335	58.949	0.92851	1.427	0.37425
ACLA	38.5	0.37518	44.509	1.00656	16.99	1.08822
Best	36.038	**QVMAX**	44.509	**ACLA**	16.99	**ACLA**
Worst	41.74	**QV**	58.949	**QVMAX2**	1.427	**QVMAX2**

Table G.35 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 3.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	26.462	0.51805	63.711	1.69738	9.826	1.66647
QV2	37.707	0.64242	41.322	0.67737	20.97	1.14925
QVMAX	41.421	0.24668	52.507	0.69876	6.071	0.54574
QVMAX2	37.491	0.34266	61.582	0.52729	0.927	0.3445
ACLA	37.843	0.49533	41.23	0.71399	20.926	0.85186
Best	26.462	**QV**	41.23	**ACLA**	20.97	**QV2**
Worst	41.421	**QVMAX**	63.711	**QV**	0.927	**QVMAX2**

Table G.36 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 3.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	34.7	0.33595	40.94	0.39772	24.359	0.50794
QV2	42.127	0.4272	55.803	0.77988	2.07	0.65205
QVMAX	23.918	0.75963	70.39	1.12874	5.691	0.64139
QVMAX2	37.979	0.32393	60.773	0.47527	1.246	0.59944
ACLA	39.904	0.36081	51.499	0.82577	8.596	0.81584
Best	23.918	**QVMAX**	40.94	**QV**	24.359	**QV**
Worst	42.127	**QV2**	70.39	**QVMAX**	1.246	**QVMAX2**

Table G.37 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 4.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	31.25	0.5469	49.505	1.01893	19.244	0.59952
QV2	38.476	0.47983	60.235	0.71603	1.288	0.42496
QVMAX	34.494	0.37525	41.051	0.48401	24.454	0.42554
QVMAX2	22.214	0.29005	76.379	0.5354	1.406	0.35208
ACLA	34.921	0.57833	35.725	0.52725	29.352	0.48329
Best	22.214	**QVMAX2**	35.725	**ACLA**	29.352	**ACLA**
Worst	38.476	**QV2**	76.379	**QVMAX2**	1.288	**QV2**

Table G.38 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 4.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	42.179	0.25624	43.168	0.55013	14.653	0.49405
QV2	38.481	0.40669	46.35	0.97021	15.168	0.79737
QVMAX	27.351	0.26972	55.397	0.6776	17.251	0.74129
QVMAX2	36.05	0.21904	62.539	0.56551	1.41	0.4868
ACLA	35.344	0.22615	36.922	0.55425	27.733	0.621
Best	27.351	**QVMAX**	36.922	**ACLA**	27.733	**ACLA**
Worst	42.179	**QV**	62.539	**QVMAX2**	1.41	**QVMAX2**

Table G.39 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 5.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	40.603	0.35911	50.974	0.46933	8.422	0.64829
QV2	36.335	0.23825	35.389	0.34537	28.275	0.45456
QVMAX	39.244	0.42048	42.847	0.72516	17.908	0.65829
QVMAX2	35.641	0.2894	62.492	0.56031	1.865	0.3815
ACLA	38.566	0.24227	37.15	0.40488	24.283	0.40871
Best	35.641	**QVMAX2**	35.389	**QV2**	28.275	**QV2**
Worst	40.603	**QV**	62.492	**QVMAX2**	1.865	**QVMAX2**

Table G.40 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 5.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Constant Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	36.296	0.37541	40.606	0.56902	23.097	0.83294
QV2	37.095	0.21702	47.61	0.64976	15.294	0.7355
QVMAX	38.16	0.47931	42.672	0.58993	19.167	0.8543
QVMAX2	37.493	0.3954	45.417	0.45629	17.089	0.44858
ACLA	36.728	0.54741	40.511	0.52134	22.761	0.76884
Best	36.296	**QV**	40.511	**ACLA**	23.097	**QV**
Worst	38.16	**QVMAX**	47.61	**QV2**	15.294	**QV2**

Table G.41 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 2.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	43.358	0.24634	55.064	0.54261	1.577	0.44956
QV2	42.11	0.39228	55.55	0.50212	2.339	0.26528
QVMAX	40.872	0.35635	58.635	0.4812	0.492	0.19386
QVMAX2	38.676	0.34113	60.418	0.52931	0.905	0.23759
ACLA	40.555	0.29957	58.311	0.46901	1.134	0.40598
Best	38.676	**QVMAX2**	55.064	**QV**	2.339	**QV2**
Worst	43.358	**QV**	60.418	**QVMAX2**	0.492	**QVMAX**

Table G.42 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 2.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	40.23	0.22256	55.769	0.75964	3.999	0.78076
QV2	39.222	0.62052	47.189	0.7741	13.588	0.4949
QVMAX	40.579	0.15746	58.052	0.26897	1.368	0.24641
QVMAX2	39.396	0.49841	59.227	0.67894	1.376	0.26291
ACLA	38.109	0.43076	52.155	1.01791	9.735	0.79418
Best	38.109	**ACLA**	47.189	**QV2**	13.588	**QV2**
Worst	40.579	**QVMAX**	59.227	**QVMAX2**	1.368	**QVMAX**

Table G.43 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 3.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	40.313	0.22716	51.65	0.53745	8.036	0.626
QV2	37.53	0.21193	55.839	0.58517	6.63	0.61781
QVMAX	37.948	0.49554	48.666	0.68334	13.386	0.82747
QVMAX2	38.853	0.24341	54.655	0.6064	6.49	0.61055
ACLA	39.115	0.34709	44.98	0.40914	15.904	0.51644
Best	37.53	**QV2**	44.98	**ACLA**	15.904	**ACLA**
Worst	40.313	**QV**	55.839	**QV2**	6.49	**QVMAX2**

Table G.44 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 3.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	31.129	0.57256	59.519	0.74878	9.351	0.50083
QV2	38.696	0.31709	43.527	0.58405	17.776	0.60161
QVMAX	40.301	0.22624	52.01	0.58329	7.687	0.48437
QVMAX2	35.823	0.19786	62.24	0.41566	1.936	0.41965
ACLA	37.819	0.40857	44.233	0.53741	17.947	0.46641
Best	31.129	**QV**	43.527	**QV2**	17.947	**ACLA**
Worst	40.301	**QVMAX**	62.24	**QVMAX2**	1.936	**QVMAX2**

Table G.45 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 4.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	36.657	0.36639	42.858	0.61595	20.484	0.60673
QV2	38.083	0.18803	57.286	0.56548	4.63	0.60172
QVMAX	24.07	0.41255	67.077	0.85104	8.852	0.49953
QVMAX2	35.785	0.172	62.064	0.53925	2.15	0.50024
ACLA	39.137	0.27928	49.925	0.66449	10.937	0.71985
Best	39.137	**ACLA**	42.858	**QV**	20.484	**QV**
Worst	24.07	**QVMAX**	67.077	**QVMAX**	2.15	**QVMAX2**

Table G.46 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 4.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	24.782	0.70541	60.315	0.71161	14.901	0.6683
QV2	35.058	0.21555	59.912	0.83976	5.029	0.94301
QVMAX	37.869	0.39751	42.458	0.46957	19.672	0.61015
QVMAX2	21.697	0.2803	76.95	0.528	1.352	0.52388
ACLA	35.638	0.3646	42.724	0.51509	21.637	0.59392
Best	21.697	**QVMAX2**	42.458	**QVMAX**	21.637	**ACLA**
Worst	37.869	**QVMAX**	76.95	**QVMAX2**	1.352	**QVMAX2**

Table G.47 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 5.0)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	38.372	0.43892	41.057	0.91095	20.57	1.15916
QV2	35.749	0.34502	49.108	0.59539	15.142	0.61232
QVMAX	27.158	0.34326	52.283	0.59119	20.559	0.43437
QVMAX2	32.756	0.13376	61.04	0.5688	6.203	0.6131
ACLA	36.218	0.44529	42.114	0.76746	21.667	1.06444
Best	32.756	**QVMAX2**	41.057	**QV**	21.667	**ACLA**
Worst	38.372	**QV**	61.04	**QVMAX2**	6.203	**QVMAX2**

Table G.48 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^G = 1$, $\mathcal{RW}_t^G > 0$ and $\mathcal{RW}_t^G < 0$. The GBR Requirement is evaluated based on NAUT-MMF user rates with the static windowing factor: $(\rho = 5.5)$ and the CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: Variable Bit Rate

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^G < 0$	Punish Reward STD[%] $\mathcal{RW}_t^G < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^G > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^G > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	38.176	0.27448	45.751	0.5773	16.071	0.58271
QV2	36.808	0.34631	52.957	0.67368	10.234	0.62461
QVMAX	36.75	0.21842	49.047	0.63246	14.201	0.57041
QVMAX2	38.371	0.19998	45.869	0.50977	15.759	0.52745
ACLA	37.174	0.28461	44.27	0.69071	18.555	0.87393
Best	36.75	**QVMAX**	44.27	**ACLA**	18.555	**ACLA**
Worst	38.371	**QVMAX2**	52.957	**QV2**	10.234	**QV2**

G.4 Summary

For the infinite buffer traffic type, the proposed static scheduling rule GPF-LM provides the highest percentages of TTIs when the GBR satisfaction domain of [91,100]% is considered. Similar results are obtained when the QV, QVMAX and ACLA RL approaches are used to train the optimal rule at each TTI

for each considered windowing factor. This is explainable since the optimal policies make use of GPF-LM for the entire scheduling session. When the CBR traffic type is used, the best performances are obtained by the ACLA policies when the windowing factor belongs to $\rho = \{2.5, 3.0, 4.0, 4.5, 5.0\}$, by the QV policies for $\rho = \{3.5, 5.5\}$ and by the QV2 policy when the factor is $\rho = 2.0$. The optimum windowing factor is $\rho = 4.0$ for the CBR traffic type when performing the ACLA policy and the STD values are minimized for the entire GBR domain (Table G.13). For the VBR traffic type, the ACLA policies indicate the best results from the viewpoints of the mean percentages of TTIs with the considered GBR satisfaction domain when the windowing factor belongs to the following set of $\rho = \{3.0, 3.5, 4.5, 5.0, 5.5\}$. The optimum windowing factor from the viewpoint of STD values for the VBR traffic type is $\rho = 3.5$ being obtained when the ACLA policy is exploited as suggested by Table G.20.

The optimum windowing factor for each traffic type is determined based on the mean percentage of TTIs when the testing rewards are punishment. For the infinite buffer traffic type, the optimum windowing factor is $\rho = 5.5$ (Table G.32) since the ACLA policy minimizes the mean percentage of TTIs with punishment rewards $\overline{p}_{TTI}^{G,PSH}$ and maximizes the percentage of TTIs when the rewards are maximized $\overline{p}_{TTI}^{G,MRW}$ over the entire domain of windowing factors. For the CBR traffic type, the optimum windowing factor is $\rho = 4.0$ (Table G.37) and it is obtained when performing the ACLA scheduling policy. When performing the scheduling policy being obtained with the same ACLA RL approach, the optimum value of the windowing factor is $\rho = 4.5$ (Table G.46) due to the fact that the mean percentage of TTIs with the punishment rewards $\overline{p}_{TTI}^{G,PSH}$ is minimized through the entire domain of the windowing factor values. It is important to remind that the optimum windowing factors are based on the simulation scenario from Table 6.3, where the maximum number of active users is 120 and the number of bearers is switched at each 1000 TTIs.

Appendix H

Performance Evaluation of Sustainable Scheduling Policies Focusing on HoL Packet Delay and PDR Objectives

H.1 Appendix Outline

In the case of HoL **D**elay and **P**DR (DP) multi-objective satisfaction criterion, the performances of the proposed scheduling policies are evaluated in terms of the particular and combined objectives. This appendix is an extension of Sub-section 7.2.4 from Chapter 7 and evaluates the performance of the obtained scheduling policies for the CBR and VBR traffic types from the viewpoint of the mean percentage of TTIs when different levels of satisfaction for the particular objective and multi-objective criteria are considered. Also, the quantity of moderate, punishment and maximum rewards for the HoL delay, PDR and DP objectives is very important in order to prove the sustainability of the proposed scheduling policies. When the PDR objective is considered, the same windowing factor is used to calculate the drop rate observations at each TTI. Then, the optimum filter length needs to be determined for each traffic type in order to maximize the mean percentage of DP feasible TTIs and to minimize at the same time, the number of DP punishment rewards.

H.2 Percentages of TTIs for the DP User Satisfaction Levels Based on the CBR and VBR Traffic Types

The grade of satisfaction for the active users is evaluated in terms of the mean percentage of TTIs and STD values when the HoL delay, PDR and DP objectives are considered such as: $\overline{p}_{TTI}^{D,x\%}, \overline{p}_{TTI}^{P,x\%}, \overline{p}_{TTI}^{DP,x\%}$, where x takes the discrete levels of $x = 91\%,...,100\%$. When the percentages of TTIs of $\overline{p}_{TTI}^{P,x\%}$ and $\overline{p}_{TTI}^{DP,x\%}$ are considered, the scheduling policies are refined and evaluated based on the following domain of the windowing factors $\rho = \{5.5, 50, 100, 200, 300, 400, 500\}$. When the windowing factor is small enough, the percentage of $\overline{p}_{TTI}^{P,x\%}$ is higher due to the limited number of TTIs in which the drop rate is computed. When the windowing factor becomes larger, the performances of $\overline{p}_{TTI}^{P,x\%}$ and $\overline{p}_{TTI}^{DP,x\%}$ decrease gradually since more dropped packets are detected in the considered time window. The optimum windowing factor should be determined in conjunction with the GBR and NGMN fairness objectives. The scheduling policies make use of four scheduling rules as follows: GPF-EDF, GPF-LOG, GPF-EXP1 and GPF-EXP2. The RL approaches which are used to refine and to improve the set of scheduling policies are: QV, QV2, QVMAX, QVMAX2 and ACLA. The CQI aggregation scheme which is performed in the controller state space computation is $(Top_3, N_{CT} = 64)$ for the entire set of simulation results. From the viewpoint of the HoL packet delay, the feasible state is determined based on the delay fraction from the HoL delay requirement (Eq. 7.2) in order to evaluate which scheduling rules and policies are able to minimize the mean of HoL packet delays while the STD values are minimized. The rest of this section is organized as follows: Tables H.1 to H.21 highlight the mean percentage of $\overline{p}_{TTI}^{D,x\%}, \overline{p}_{TTI}^{P,x\%}, \overline{p}_{TTI}^{DP,x\%}$ and the STD values for the obtained policies and for the existing scheduling rules when the CBR traffic type is simulated and Tables H.22 to H.42 evaluate the scheduling policies from the perspective of the same indicators but for the VBR traffic type.

Table H.1 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / HOL Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.413	0.50052	68.298	0.96754	36.623	2.74998	9.334	1.532
92% Sat.	2.166	0.43165	67.997	0.97935	33.413	2.78887	8.423	1.32567
93% Sat.	1.954	0.39295	67.668	0.99847	30.776	2.71629	7.534	1.1725
94% Sat.	1.833	0.3963	67.163	0.91734	27.048	2.63089	6.812	1.18799
95% Sat.	0.948	0.19586	62.422	1.25615	19.549	2.50665	3.985	0.80116
96% Sat.	0.646	0.09674	60.923	1.18932	16.69	2.39855	2.982	0.70517
97% Sat.	0.49	0.08568	58.049	1.58273	12.483	1.8828	2.196	0.46694
98% Sat.	0.423	0.0964	55.77	1.88138	9.699	1.67216	1.911	0.47494
99% Sat.	0.383	0.10106	51.468	2.09016	9.064	1.54698	1.848	0.47769
100% Sat.	0.352	0.10237	51.334	2.0649	8.997	1.54459	1.781	0.47838

Policies / HOL Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	41.595	2.40377	65.06	0.67848	62.294	0.82245	69.852	0.81537	70.49	0.88907
92% Sat.	37.792	2.41896	63.684	0.70761	61.263	0.97477	69.79	0.80707	70.412	0.89374
93% Sat.	34.645	2.36103	62.85	0.70024	60.206	1.14399	69.708	0.78408	70.341	0.89595
94% Sat.	30.205	2.36148	61.272	0.87977	58.834	1.2665	69.596	0.81077	70.207	0.92218
95% Sat.	21.872	2.33656	58.815	1.0673	51.45	1.46186	68.244	0.87376	68.188	1.07583
96% Sat.	18.526	2.26196	57.14	1.34677	49.273	1.59484	67.804	0.99055	67.651	1.19609
97% Sat.	13.472	1.90619	55.62	1.46826	45.575	1.75717	67.4	1.08172	67.109	1.26949
98% Sat.	10.539	1.71036	53.933	1.61167	42.683	2.133	66.979	1.08444	66.506	1.27958
99% Sat.	9.437	1.61727	51.387	1.62631	40.894	2.34858	66.241	1.07263	65.38	1.24517
100% Sat.	9.367	1.6214	51.19	1.58313	40.684	2.37858	66.13	1.06846	65.195	1.2224

Table H.2 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with the static windowing factor of $(\rho = 5.5)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / PDR Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	88.147	0.28115	88.247	0.38571	89.35	0.20322	87.757	0.46389
92% Sat.	88.129	0.27793	88.124	0.42323	89.27	0.25485	87.596	0.55321
93% Sat.	88.106	0.27648	87.996	0.42939	89.214	0.26099	87.442	0.56975
94% Sat.	88.081	0.27909	87.84	0.46881	89.138	0.2686	87.262	0.63306
95% Sat.	87.956	0.35545	87.319	0.52406	88.913	0.44194	86.535	0.60874
96% Sat.	87.92	0.35377	86.921	0.56512	88.822	0.47015	86.022	0.68371
97% Sat.	87.86	0.3448	86.443	0.5223	88.598	0.55658	85.323	0.72177
98% Sat.	87.611	0.37034	85.561	0.69849	88.176	0.6978	84.264	0.80198
99% Sat.	87.333	0.43017	84.815	0.64812	87.508	0.7352	83.184	0.8247
100% Sat.	86.811	0.52931	83.125	0.66899	86.469	0.83178	81.288	0.86254

Policies / PDR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	88.434	0.20799	88.142	0.19025	88.584	0.33307	88.958	0.25646	87.904	0.3384
92% Sat.	88.396	0.18912	88.112	0.19101	88.487	0.40811	88.884	0.25276	87.824	0.33178
93% Sat.	88.365	0.18009	88.054	0.20016	88.389	0.40618	88.823	0.25241	87.726	0.3469
94% Sat.	88.319	0.17782	87.99	0.22141	88.297	0.45753	88.728	0.26676	87.614	0.38534
95% Sat.	88.082	0.34232	87.802	0.40727	88.002	0.54798	88.421	0.45845	87.325	0.5071
96% Sat.	88.029	0.34141	87.717	0.41394	87.767	0.5812	88.318	0.48138	87.009	0.53436
97% Sat.	87.946	0.32256	87.594	0.39718	87.37	0.55423	88.062	0.59013	86.742	0.45665
98% Sat.	87.717	0.34234	87.277	0.44983	86.708	0.71522	87.634	0.74325	86.038	0.63598
99% Sat.	87.257	0.48132	86.8	0.52045	86.004	0.65949	87.011	0.81449	85.437	0.64664
100% Sat.	86.786	0.54656	86.117	0.55472	84.427	0.74656	85.997	0.89736	83.56	0.799

Table H.3 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-objective satisfaction levels based on PDR-MMF with static windowing factor: $(\rho = 5.5)$. CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$. Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.06	0.47529	67.71	0.9709	36.148	2.73209	8.929	1.51111
92% Sat.	1.844	0.40507	67.399	0.98644	32.95	2.76864	8.032	1.30127
93% Sat.	1.647	0.37634	67.059	1.00615	30.32	2.70011	7.146	1.15308
94% Sat.	1.554	0.37875	66.546	0.92557	26.603	2.61131	6.446	1.17003
95% Sat.	0.698	0.1776	61.794	1.26117	19.11	2.48585	3.626	0.77428
96% Sat.	0.415	0.08416	60.288	1.18849	16.261	2.37556	2.643	0.67942
97% Sat.	0.326	0.08476	57.403	1.57548	12.083	1.85807	1.904	0.45422
98% Sat.	0.279	0.0921	55.101	1.87656	9.309	1.64519	1.635	0.46048
99% Sat.	0.27	0.0929	50.826	2.08833	8.718	1.52123	1.614	0.46163
100% Sat.	0.26	0.09481	50.709	2.06704	8.685	1.52393	1.602	0.46746

Policies / DP Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	41.09	2.39182	64.416	0.69298	61.672	0.81956	69.31	0.83217	69.843	0.89792
92% Sat.	37.302	2.40434	63.034	0.72256	60.617	0.96648	69.238	0.81834	69.755	0.90544
93% Sat.	34.16	2.34227	62.192	0.71271	59.548	1.13865	69.146	0.79638	69.675	0.9049
94% Sat.	29.738	2.33922	60.605	0.89124	58.159	1.25716	69.024	0.8235	69.522	0.93513
95% Sat.	21.42	2.31492	58.147	1.07516	50.758	1.44965	67.661	0.88058	67.484	1.09534
96% Sat.	18.085	2.23404	56.457	1.35685	48.568	1.57385	67.206	0.99664	66.937	1.21819
97% Sat.	13.064	1.87564	54.937	1.47627	44.843	1.7346	66.785	1.09237	66.369	1.28281
98% Sat.	10.144	1.68125	53.207	1.6071	41.905	2.11343	66.329	1.09129	65.714	1.29176
99% Sat.	9.08	1.58695	50.687	1.61082	40.115	2.34189	65.59	1.08438	64.581	1.264
100% Sat.	9.051	1.59149	50.474	1.56354	39.894	2.37562	65.455	1.07427	64.369	1.23422

Table H.4 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.512	0.74137	68.086	0.44238	38.127	3.30427	9.901	1.93854
92% Sat.	2.208	0.70398	67.763	0.49302	35.063	3.07225	8.782	1.93548
93% Sat.	2.039	0.68662	67.49	0.47898	32.251	3.14947	7.993	1.88294
94% Sat.	1.899	0.66471	66.904	0.56089	28.63	2.98478	7.328	1.75701
95% Sat.	0.898	0.19792	62.865	0.93065	20.733	2.45967	3.935	0.86279
96% Sat.	0.687	0.16146	61.365	1.1534	17.63	2.32352	3.076	0.83169
97% Sat.	0.531	0.14297	59.26	0.93919	13.458	2.13682	2.259	0.74356
98% Sat.	0.454	0.13732	57.112	1.11206	10.684	2.07892	1.916	0.64298
99% Sat.	0.409	0.13485	53.046	0.97503	10.012	1.90142	1.857	0.64067
100% Sat.	0.37	0.14206	52.893	1.00011	9.954	1.89693	1.801	0.63596

Policies / DP Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	60.387	1.13794	69.559	0.28029	44.421	2.99688	49.936	1.02483	65.521	0.80895
92% Sat.	58.753	1.01964	69.369	0.29193	41.709	2.82955	49.664	0.9915	65.359	0.82604
93% Sat.	57.762	0.98593	69.256	0.28226	39.244	2.93339	49.429	0.97465	65.196	0.82691
94% Sat.	55.362	0.84699	68.672	0.33813	35.831	2.81089	49.166	0.91897	65.045	0.82135
95% Sat.	50.331	1.21102	66.368	0.58165	27.991	2.46793	47.767	0.88618	63.729	0.96642
96% Sat.	47.639	1.2808	65.116	0.58911	24.914	2.44196	47.32	0.91126	63.222	0.96181
97% Sat.	43.936	1.05172	63.155	0.84291	21.008	2.27592	46.868	0.90217	62.596	1.02609
98% Sat.	39.469	1.17156	61.534	1.03738	18.083	2.18485	46.355	0.73976	62.203	1.107
99% Sat.	37	1.02997	57.233	1.16782	17.331	1.98583	46.078	0.79181	60.842	1.13835
100% Sat.	36.854	1.03902	57.067	1.18271	17.125	2.00168	45.996	0.79054	60.81	1.14051

Table H.5 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR.

Policies / PDR Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	79.418	0.77267	79.789	0.88951	81.325	0.7035	78.618	0.72591
92% Sat.	79.375	0.76548	79.576	0.92997	81.188	0.68009	78.319	0.81088
93% Sat.	79.321	0.75727	79.289	1.01592	81.061	0.72459	78.038	0.86242
94% Sat.	79.148	0.76477	78.725	1.05545	80.766	0.72313	77.364	0.90443
95% Sat.	78.992	0.78016	77.99	0.89979	80.384	0.70439	76.281	0.91392
96% Sat.	78.759	0.76647	77.231	1.00263	79.864	0.80393	75.312	0.94153
97% Sat.	78.446	0.80314	76.393	1.06067	79.222	0.92014	74.265	1.09953
98% Sat.	77.697	0.78315	74.85	1.26506	78.22	0.98743	72.361	1.49013
99% Sat.	77.322	0.96864	73.751	1.1086	77.439	0.91305	70.634	1.50467
100% Sat.	76.552	1.1117	70.904	1.00445	76.21	1.10496	67.832	1.11627

Policies / PDR Satisfaction	QV [Mean]	QV [STD]	QV2 [Mean]	QV2 [STD]	QVMAX [Mean]	QVMAX [STD]	QVMAX2 [Mean]	QVMAX2 [STD]	ACLA- [Mean]	ACLA [STD]
91% Sat.	80.23	1.03873	80.49	0.82552	79.922	0.71508	79.686	0.63875	79.473	0.59885
92% Sat.	80.036	1.00259	80.445	0.82682	79.68	0.7183	79.423	0.69155	79.377	0.62178
93% Sat.	79.787	0.98968	80.384	0.83048	79.473	0.77523	79.21	0.74215	79.029	0.74741
94% Sat.	79.187	1.05923	80.216	0.8268	78.96	0.80931	78.677	0.85574	78.54	0.77433
95% Sat.	78.531	1.01502	80.062	0.78717	78.552	0.8579	78.081	0.84704	78.142	0.77371
96% Sat.	77.84	1.1225	79.65	0.853	77.978	0.92832	77.398	0.91319	77.539	0.84374
97% Sat.	76.993	1.08171	79.216	0.93664	77.303	1.01284	76.767	1.02115	77.024	0.92507
98% Sat.	75.484	1.23573	78.415	0.93641	75.914	1.11367	75.299	1.18516	75.573	0.99439
99% Sat.	74.461	1.10078	77.915	0.91832	74.745	1.01485	73.735	1.18023	74.756	1.09141
100% Sat.	71.731	0.96255	77.097	0.98235	72.289	0.73151	71.162	0.8355	72.421	0.77978

Table H.6 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	1.936	0.72731	63.813	0.41224	36.19	3.19193	9.047	1.89436
92% Sat.	1.673	0.69766	63.524	0.50046	33.317	2.98378	7.981	1.88896
93% Sat.	1.558	0.67951	63.311	0.50844	30.706	3.05835	7.262	1.84111
94% Sat.	1.462	0.65585	62.566	0.541	27.232	2.90244	6.649	1.71551
95% Sat.	0.534	0.18872	58.574	0.83989	19.484	2.38444	3.344	0.8454
96% Sat.	0.374	0.14928	57.174	1.07748	16.507	2.25901	2.555	0.81367
97% Sat.	0.283	0.13332	55.271	0.90969	12.628	2.09218	1.83	0.73237
98% Sat.	0.231	0.1264	53.047	1.06482	9.882	2.02467	1.523	0.63012
99% Sat.	0.225	0.12716	49.589	1.03002	9.355	1.86005	1.515	0.63029
100% Sat.	0.222	0.12728	49.556	1.05199	9.35	1.85879	1.511	0.63037

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	55.657	1.12085	64.23	0.32241	39.48	2.99539	46.879	0.99024	63.214	0.68356
92% Sat.	54.068	1.01166	64.057	0.33785	36.726	2.83475	46.628	0.95102	63.099	0.69919
93% Sat.	53.151	0.99322	63.948	0.32846	34.268	2.93545	46.464	0.93826	63.007	0.672
94% Sat.	50.577	0.809	63.282	0.33295	30.665	2.74665	46.224	0.87699	62.725	0.69708
95% Sat.	45.619	1.17409	61.123	0.64961	22.877	2.3728	44.978	0.85828	61.473	0.87654
96% Sat.	43.003	1.20476	59.767	0.6595	19.782	2.30048	44.55	0.87863	60.946	0.89339
97% Sat.	39.552	0.97745	57.751	0.92161	15.901	2.18966	44.14	0.86904	60.433	0.97013
98% Sat.	35.011	1.12208	56.071	1.08223	12.871	2.08107	43.622	0.7048	59.814	1.0064
99% Sat.	33.122	0.9396	51.82	1.17805	12.234	1.91465	43.393	0.7654	58.5	0.99302
100% Sat.	33.116	0.93815	51.813	1.17773	12.228	1.91443	43.388	0.76631	58.495	0.99234

Table H.7 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.325	0.51057	68.194	0.85626	34.646	2.22551	8.898	1.18995
92% Sat.	2.039	0.4064	67.879	0.91968	31.733	2.30596	7.937	1.03603
93% Sat.	1.882	0.40232	67.54	0.91524	29.173	2.26838	7.154	0.94274
94% Sat.	1.774	0.41558	67.048	0.99565	25.964	2.05178	6.531	0.88996
95% Sat.	0.9	0.19863	62.954	1.1412	18.577	2.07105	3.655	0.61358
96% Sat.	0.649	0.13736	61.304	1.17258	15.813	1.74215	2.857	0.64166
97% Sat.	0.485	0.09816	58.396	1.37961	12.076	1.64077	2.091	0.45839
98% Sat.	0.411	0.09745	56.325	1.72792	9.539	1.24931	1.78	0.40855
99% Sat.	0.373	0.10158	51.707	1.59758	9.022	1.15223	1.715	0.41207
100% Sat.	0.339	0.10346	51.558	1.54858	8.96	1.15099	1.664	0.41707

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	38.34	1.65654	69.204	0.70142	64.824	1.51504	62.656	0.97656	67.674	0.72117
92% Sat.	36.425	1.70996	69.104	0.71277	64.71	1.53797	62.311	1.01395	67.476	0.72548
93% Sat.	34.534	1.75822	68.994	0.70219	64.525	1.53993	61.978	1.01061	67.149	0.72342
94% Sat.	32.041	1.63419	68.813	0.74351	64.322	1.67243	61.557	1.12248	67.01	0.73119
95% Sat.	24.889	1.54809	65.148	0.92335	61.714	1.77324	59.444	1.12085	65.534	0.7962
96% Sat.	22.464	1.39313	64.009	1.01783	61.092	1.84834	58.347	1.16932	65.007	0.83966
97% Sat.	18.949	1.30572	62.522	1.1277	60.4	1.8725	56.687	1.3604	64.396	0.86852
98% Sat.	16.387	1.02605	61.517	1.47921	59.853	1.97272	55.407	1.70225	63.969	1.04316
99% Sat.	15.184	1.03923	60.525	1.63549	59.053	1.77072	52.306	1.60096	63.077	1.2178
100% Sat.	15.108	1.02799	60.336	1.5933	58.943	1.76202	52.252	1.59675	62.937	1.20938

Table H.8 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	70.183	1.1185	71.765	1.15372	73.291	0.98395	69.549	1.18737
92% Sat.	70.093	1.13235	71.218	1.38149	73.113	1.01246	68.973	1.24291
93% Sat.	69.936	1.0877	70.586	1.36497	72.895	1.02812	68.358	1.21514
94% Sat.	69.506	1.1847	69.641	1.38288	72.29	1.03188	67.255	1.19197
95% Sat.	69.068	1.38552	68.426	1.46738	71.474	1.31573	65.904	1.3258
96% Sat.	68.398	1.4378	67.123	1.58999	70.601	1.35263	64.593	1.5058
97% Sat.	68.045	1.42154	66.527	1.50244	70.154	1.2852	63.79	1.71618
98% Sat.	66.092	1.61573	64.215	1.59988	67.967	1.36564	61.157	2.05306
99% Sat.	65.853	1.66829	63.034	1.86018	67.265	1.55673	59.42	2.21643
100% Sat.	64.951	1.73923	60.786	1.74225	65.848	1.43934	57.107	2.17921

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	70.945	1.14402	72.008	1.06653	72.466	1.09464	73.263	1.05755	72.39	1.07899
92% Sat.	70.257	1.20665	71.621	1.09918	72.26	1.06094	73.068	1.05258	72.13	1.05042
93% Sat.	69.58	1.11835	71.107	1.04195	72.094	0.98791	72.924	1.01532	71.877	0.97945
94% Sat.	68.385	1.25004	70.33	1.05425	71.65	1.01441	72.447	1.03249	71.348	1.0483
95% Sat.	67.533	1.46471	69.474	1.16815	71.196	1.25828	71.725	1.2924	70.844	1.31022
96% Sat.	66.269	1.50146	68.429	1.27414	70.488	1.23675	70.934	1.30585	70.05	1.37948
97% Sat.	65.786	1.47726	67.842	1.22184	70.112	1.2553	70.487	1.28353	69.715	1.38059
98% Sat.	63.711	1.37204	65.562	1.27646	68.398	1.39743	68.749	1.46119	67.87	1.77069
99% Sat.	62.617	1.71146	64.479	1.53923	67.492	1.4452	68.128	1.59271	67.365	1.86585
100% Sat.	60.483	1.59773	62.438	1.27337	65.768	1.37792	67.166	1.65598	66.233	1.7391

Table H.9 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	1.633	0.4544	59.053	1.14953	30.81	1.86701	7.624	1.05565
92% Sat.	1.407	0.35643	58.533	1.33257	28.302	1.8962	6.728	0.92049
93% Sat.	1.31	0.35867	58.168	1.30367	26.265	1.93343	6.091	0.83132
94% Sat.	1.242	0.37112	57.236	1.34385	23.368	1.71814	5.549	0.78719
95% Sat.	0.503	0.19144	53.303	1.36403	16.445	1.73909	2.903	0.55514
96% Sat.	0.324	0.12882	51.698	1.31408	13.973	1.48822	2.208	0.57037
97% Sat.	0.245	0.09547	49.359	1.37505	10.801	1.49706	1.579	0.41344
98% Sat.	0.204	0.08401	47.016	1.70696	8.305	1.11382	1.306	0.35618
99% Sat.	0.201	0.08373	43.648	1.75083	8	1.04615	1.3	0.35566
100% Sat.	0.192	0.08823	43.614	1.71696	7.995	1.04792	1.298	0.35673

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	30.781	1.66021	59.566	1.0708	56.357	1.75378	58.384	1.1496	60.145	0.98776
92% Sat.	28.524	1.649	59.263	1.1434	56.299	1.75467	58.276	1.17813	59.915	0.9567
93% Sat.	26.49	1.65756	58.994	1.06812	56.209	1.7624	58.106	1.13986	59.725	0.8881
94% Sat.	23.547	1.54278	58.334	1.06902	56.047	1.90109	57.906	1.29069	59.269	0.96142
95% Sat.	16.701	1.51202	54.565	1.23067	53.764	1.93846	56.107	1.26499	58.079	1.09198
96% Sat.	14.323	1.27726	53.29	1.23858	53.033	1.96961	54.941	1.24684	57.318	1.09823
97% Sat.	11.11	1.19658	51.555	1.20894	52.613	1.99019	53.659	1.40673	56.981	1.12693
98% Sat.	8.662	0.91531	49.893	1.58303	51.96	2.02471	52.074	1.78806	56.167	1.3589
99% Sat.	8.392	0.89891	48.841	1.64267	51.368	1.82005	49.141	1.70735	55.386	1.40159
100% Sat.	8.379	0.89626	48.815	1.62597	51.355	1.8106	49.121	1.69852	55.376	1.40315

Table H.10 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.619	0.6636	68.188	0.65557	36.59	2.92585	9.43	1.40777
92% Sat.	2.354	0.6634	67.848	0.65607	33.392	2.78007	8.482	1.46141
93% Sat.	2.166	0.59736	67.499	0.63186	30.654	2.81881	7.778	1.31638
94% Sat.	2.048	0.5976	67.064	0.69698	27.121	2.6039	7.06	1.20652
95% Sat.	1.006	0.32501	62.667	1.09336	19.652	2.25432	4.007	0.87183
96% Sat.	0.699	0.20896	60.936	1.28902	16.509	2.06687	2.988	0.6496
97% Sat.	0.517	0.16013	58.075	1.53045	12.746	1.6436	2.341	0.53918
98% Sat.	0.455	0.15937	55.986	1.64574	9.999	1.48394	2.046	0.51082
99% Sat.	0.415	0.15184	52.409	2.03152	9.196	1.31295	1.977	0.51157
100% Sat.	0.389	0.14506	52.282	2.03293	9.129	1.31044	1.926	0.50973

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	47.474	1.88856	48.621	1.15706	49.37	2.12736	69.297	0.63062	61.967	1.53424
92% Sat.	46.228	1.85972	48.329	1.12948	46.858	2.05382	69.096	0.61748	61.889	1.51812
93% Sat.	44.989	1.88898	47.923	1.11225	45.03	2.06428	68.969	0.60109	61.834	1.53032
94% Sat.	43.552	1.92078	47.664	1.1861	42.052	1.98989	68.671	0.58454	61.727	1.51993
95% Sat.	37.472	1.51343	45.801	1.26293	36.748	2.07221	64.926	0.99297	60.229	1.68344
96% Sat.	35.254	1.52804	45.194	1.33804	34.058	2.35382	63.671	1.22433	59.72	1.74988
97% Sat.	32.128	1.38536	44.665	1.37291	30.806	1.94613	62.224	1.39963	59.407	1.78251
98% Sat.	29.209	1.35743	44.228	1.54702	28.027	1.78255	60.439	1.38846	59.101	1.77906
99% Sat.	28.005	1.65073	43.748	1.62023	26.735	1.78474	59.34	1.40806	58.602	1.70099
100% Sat.	27.812	1.6462	43.71	1.62909	26.662	1.79338	59.146	1.42152	58.538	1.70017

Table H.11 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	55.575	1.2068	57.863	1.12135	59.97	0.7294	55.353	0.88531
92% Sat.	55.342	1.17189	57.242	1.40159	59.499	0.9564	54.645	1.27223
93% Sat.	55.101	1.18129	56.603	1.4876	58.994	1.05772	54.047	1.44749
94% Sat.	54.026	1.29007	55.025	1.42799	57.788	0.96966	51.943	1.59495
95% Sat.	53.588	1.35881	53.584	1.46543	56.866	1.07066	50.088	1.60166
96% Sat.	52.558	1.66599	51.3	1.62487	55.309	1.17248	47.755	1.65262
97% Sat.	51.826	1.71162	50.054	1.68356	54.137	1.02431	46.447	1.42546
98% Sat.	48.749	2.10615	47.017	1.10729	50.548	0.69132	43.071	1.55035
99% Sat.	48.264	2.07684	45.526	0.97451	49.537	1.06269	41.824	1.17228
100% Sat.	46.726	1.98929	44.21	0.92392	47.909	0.91446	40.618	1.59517

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	57.591	1.13613	57.121	0.99989	57.439	0.96996	59.071	0.7774	57.063	0.96895
92% Sat.	57.033	1.36599	56.644	1.25163	56.841	1.29204	58.811	0.85142	56.585	1.22263
93% Sat.	56.433	1.63079	56.126	1.32188	56.266	1.45119	58.462	0.8952	56.058	1.28297
94% Sat.	55.192	1.63591	54.52	1.33398	54.865	1.44032	57.602	0.92529	54.521	1.38803
95% Sat.	54.237	1.5969	53.266	1.36964	53.669	1.32038	57.116	0.9099	53.18	1.44085
96% Sat.	53.07	1.77876	51.334	1.42772	51.84	1.65526	55.731	0.92219	51.22	1.65015
97% Sat.	51.993	1.93143	50.025	1.73352	50.742	1.81547	54.891	0.79313	49.955	1.78903
98% Sat.	49.029	1.63779	46.818	1.1358	47.458	1.2299	51.351	0.80665	46.681	1.21368
99% Sat.	47.912	1.68628	45.234	0.97675	46.258	1.36763	50.629	0.88029	45.378	1.00358
100% Sat.	46.576	1.51442	43.788	1.10252	45.25	1.45311	48.862	0.68849	44.048	1.07632

Table H.12 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	1.717	0.50422	48.829	0.83165	27.826	2.07653	7.14	1.11168
92% Sat.	1.532	0.51377	48.322	0.96052	25.483	1.93435	6.382	1.16148
93% Sat.	1.433	0.46118	47.925	0.92656	23.902	1.95278	5.951	1.06638
94% Sat.	1.36	0.4545	46.669	1.106	21.053	1.79078	5.341	0.93717
95% Sat.	0.561	0.23922	42.643	1.54949	14.821	1.55143	2.858	0.68521
96% Sat.	0.346	0.1489	40.729	1.73486	12.358	1.53793	2.034	0.50539
97% Sat.	0.26	0.11024	38.637	2.01161	9.988	1.3328	1.602	0.42232
98% Sat.	0.228	0.10775	35.723	2.01967	7.267	1.16245	1.346	0.39274
99% Sat.	0.222	0.1077	34.204	1.62987	7.026	1.18008	1.339	0.39082
100% Sat.	0.219	0.10848	34.191	1.63825	7.021	1.18074	1.336	0.39243

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	36.706	1.8271	42.178	1.05725	34.192	1.77093	48.309	0.53714	46.687	0.80927
92% Sat.	35.289	1.66829	42.071	1.03711	31.868	1.65483	47.975	0.54639	46.418	0.75483
93% Sat.	34.07	1.60802	41.999	1.02918	30.345	1.65363	47.738	0.44988	46.214	0.70615
94% Sat.	31.861	1.77417	41.813	1.11048	27.605	1.51961	46.934	0.52295	45.19	0.89641
95% Sat.	25.964	1.50003	40.552	1.2447	23.114	1.65212	43.464	0.89587	43.482	1.16786
96% Sat.	23.566	1.77588	40.113	1.27496	20.716	1.85521	41.569	1.05267	42.363	1.05668
97% Sat.	20.604	1.59416	39.851	1.30079	18.306	1.59493	39.964	1.16361	41.62	1.19702
98% Sat.	16.989	1.7056	39.4	1.47249	15.534	1.44638	37.167	1.35325	39.947	1.63389
99% Sat.	15.906	1.60696	39.208	1.46074	15.235	1.4655	36.317	1.46743	39.193	1.46972
100% Sat.	15.902	1.60723	39.205	1.46123	15.229	1.46761	36.304	1.47048	39.19	1.47026

Table H.13 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.36	0.56537	68.38	1.08567	36.623	4.2264	8.96	2.26413
92% Sat.	2.095	0.55235	68.08	1.08716	33.531	4.22127	8.076	2.10116
93% Sat.	1.968	0.52798	67.679	0.95333	30.635	4.21197	7.36	1.98489
94% Sat.	1.829	0.51928	67.297	0.94536	27.148	3.97014	6.606	1.83911
95% Sat.	0.953	0.26971	62.502	1.64942	19.587	3.33566	3.884	1.09282
96% Sat.	0.641	0.14724	61.201	1.7413	16.477	2.99051	2.867	0.77868
97% Sat.	0.481	0.13093	58.629	2.01085	12.774	2.64923	2.163	0.68452
98% Sat.	0.411	0.1185	56.83	2.47508	9.86	2.29012	1.804	0.58346
99% Sat.	0.374	0.11658	52.727	2.34055	9.135	2.08114	1.736	0.58153
100% Sat.	0.344	0.11346	52.598	2.33548	9.072	2.08601	1.685	0.57692

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	62.064	2.86686	37.677	1.2184	45.627	3.79106	10.369	2.37923	11.461	2.40693
92% Sat.	61.988	2.86448	37.13	1.16741	43.375	3.83722	9.368	2.18791	10.369	2.19772
93% Sat.	61.915	2.86909	36.525	1.16844	40.961	3.82338	8.496	2.042	9.47	2.07736
94% Sat.	61.791	2.91559	35.832	1.13518	38.137	3.72599	7.581	1.90255	8.365	2.04753
95% Sat.	59.505	2.87721	33.52	1.21886	30.803	3.25858	4.48	1.13579	4.979	1.31338
96% Sat.	59.167	2.92937	32.789	1.13236	27.948	2.95301	3.375	0.86529	3.638	1.00453
97% Sat.	58.728	3.044	32.106	1.14765	24.497	2.81116	2.57	0.75662	2.835	0.87603
98% Sat.	58.475	3.0768	31.308	1.0664	22.004	2.52732	2.124	0.64404	2.35	0.79786
99% Sat.	57.646	3.02089	30.876	1.18618	20.933	2.22344	2.047	0.6375	2.267	0.78591
100% Sat.	57.482	3.04189	30.853	1.18712	20.703	2.24518	1.984	0.63055	2.206	0.77949

Table H.14 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	42.354	2.51348	45.766	1.90709	48.172	1.78075	41.837	1.95771
92% Sat.	42.058	2.58339	44.787	2.08803	47.539	2.16806	40.559	2.40105
93% Sat.	41.746	2.49286	43.941	2.15283	47.279	2.2537	39.864	2.35631
94% Sat.	40.5	2.70787	41.948	2.21506	45.738	2.36978	38.039	2.2302
95% Sat.	39.751	2.9234	39.526	2.34606	44.277	2.62829	35.511	2.06964
96% Sat.	37.912	3.33966	37.743	2.10875	42.038	2.89448	33.882	2.07615
97% Sat.	37.432	3.38186	36.52	2.27263	41.193	2.87106	32.82	2.25734
98% Sat.	34.718	3.62934	33.507	1.99851	38.09	3.1583	30.542	2.17673
99% Sat.	34.292	3.63279	32.887	1.88891	37.548	3.09887	29.907	2.03987
100% Sat.	34.038	3.5313	32.375	1.67793	37.244	3.01736	29.564	1.83284

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	44.177	2.15695	43.801	2.02939	46.885	2.70642	43.234	2.61556	44.761	2.18281
92% Sat.	43.295	2.27634	42.659	2.10921	46.608	2.79305	42.858	2.62156	44.315	2.19365
93% Sat.	42.682	2.44098	41.934	2.18326	46.281	2.77642	42.358	2.66743	43.887	2.44139
94% Sat.	40.968	2.42295	40.094	2.13773	45.276	2.66938	41.347	2.88041	42.148	2.43559
95% Sat.	38.721	2.40667	37.963	2.23145	44.566	2.85401	39.92	3.25134	40.284	2.34223
96% Sat.	36.946	2.04996	36.445	2.03317	42.64	3.04732	37.903	3.54919	38.202	2.51233
97% Sat.	36.365	2.02385	35.678	1.93761	42.051	2.99457	37.068	3.48852	37.221	2.63001
98% Sat.	33.568	2.24929	32.913	2.03576	39.11	3.25352	34.294	3.40258	34.072	2.6622
99% Sat.	32.897	2.16355	32.405	1.84749	38.675	3.20492	33.88	3.28981	33.558	2.53384
100% Sat.	32.428	2.06456	32.21	1.87458	38.288	3.15368	33.348	3.36715	33.116	2.34832

Table H.15 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor: $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	1.222	0.37932	40.259	1.54109	23.696	3.23149	5.647	1.8088
92% Sat.	1.074	0.37417	39.347	1.62495	21.653	3.18911	5.037	1.66893
93% Sat.	1.042	0.36703	38.879	1.59272	20.789	3.19443	4.802	1.63548
94% Sat.	0.965	0.34969	37.305	1.81062	18.212	2.9559	4.16	1.47624
95% Sat.	0.465	0.19931	32.702	2.38797	12.517	2.39278	2.325	0.90401
96% Sat.	0.264	0.0848	30.9	2.32943	10.26	2.2152	1.577	0.57619
97% Sat.	0.199	0.08198	29.273	2.2557	8.597	2.12312	1.23	0.53754
98% Sat.	0.149	0.07321	25.939	2.11231	5.802	1.69708	0.934	0.41648
99% Sat.	0.147	0.07243	25.449	1.95358	5.737	1.66988	0.931	0.41517
100% Sat.	0.146	0.07251	25.441	1.95129	5.732	1.66913	0.929	0.41525

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	32.31	2.46603	29.575	1.14865	27.615	2.92892	6.28	1.83811	7.522	1.85984
92% Sat.	31.672	2.182	29.324	1.18006	26.229	2.93447	5.659	1.6462	6.781	1.66653
93% Sat.	31.465	2.24003	29.246	1.20039	25.44	2.90352	5.374	1.61314	6.464	1.65402
94% Sat.	30.281	2.55864	28.748	1.13731	23.427	2.83966	4.666	1.50456	5.512	1.57922
95% Sat.	27.233	2.51583	27.326	1.30604	18.355	2.34978	2.761	0.96631	2.994	1.04023
96% Sat.	26.259	2.40742	26.956	1.25071	16.052	2.1396	1.869	0.62322	1.981	0.72371
97% Sat.	25.87	2.37657	26.695	1.29188	14.463	2.14457	1.459	0.57408	1.581	0.6354
98% Sat.	23.535	2.10317	25.917	1.2295	11.831	1.86869	1.092	0.46598	1.159	0.51647
99% Sat.	23.098	1.84044	25.901	1.22971	11.44	1.74836	1.088	0.46466	1.155	0.51602
100% Sat.	23.095	1.8403	25.897	1.22936	11.426	1.73799	1.085	0.46446	1.153	0.51566

Table H.16. Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.549	0.70215	67.887	0.74226	34.246	2.73231	8.84	1.93186
92% Sat.	2.325	0.66446	67.634	0.72998	31.223	2.6587	8.03	1.90027
93% Sat.	2.068	0.53041	67.233	0.62395	28.707	2.81968	7.201	1.59681
94% Sat.	1.942	0.51858	66.711	0.54289	25.367	2.73295	6.698	1.53349
95% Sat.	0.956	0.29038	61.806	0.96116	18.046	2.75605	3.854	1.25827
96% Sat.	0.708	0.21817	60.265	1.07126	15.018	2.60348	2.944	0.99044
97% Sat.	0.532	0.14209	57.576	1.62915	11.403	2.2356	2.135	0.70239
98% Sat.	0.462	0.14306	55.052	2.12605	9.06	2.0085	1.873	0.65777
99% Sat.	0.424	0.15038	50.991	2.35248	8.624	1.97441	1.817	0.66532
100% Sat.	0.388	0.14909	50.882	2.34489	8.553	1.98026	1.768	0.66413

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	50.851	1.51855	55.57	1.26758	70.604	0.76467	70.259	0.70581	65.565	0.34006
92% Sat.	48.922	1.45118	54.862	1.33158	70.526	0.75404	70.196	0.70185	64.742	0.3001
93% Sat.	47.046	1.50218	54.292	1.37892	70.388	0.69581	70.113	0.68669	64.04	0.31467
94% Sat.	44.672	1.56201	53.81	1.33537	70.255	0.68223	69.97	0.69881	63.264	0.26542
95% Sat.	36.285	1.54989	50.318	1.4046	67.881	0.81196	68.103	0.70141	58.226	0.52337
96% Sat.	32.579	1.57049	49.191	1.41831	67.254	0.81245	67.629	0.69073	55.977	0.72703
97% Sat.	27.352	1.47032	47.853	1.17248	66.604	0.95359	67.149	0.8068	54.661	0.97936
98% Sat.	22.785	1.6078	47.281	1.16111	65.916	1.02331	66.735	0.77526	53.15	1.31099
99% Sat.	20.922	1.6575	46.684	1.30303	64.701	1.33446	65.9	1.06651	50.748	1.59292
100% Sat.	20.808	1.66174	46.599	1.29544	64.502	1.32425	65.713	1.04437	50.711	1.59286

Table H.17 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	32.792	2.69653	36.482	2.94585	38.216	2.25676	31.371	2.85416
92% Sat.	32.17	2.6524	35.477	2.45566	37.503	2.12473	30.604	2.66323
93% Sat.	31.909	2.8053	34.82	2.28665	36.991	2.27662	30.204	2.53936
94% Sat.	31.04	3.20354	33.858	2.4343	35.764	2.76452	28.267	2.48015
95% Sat.	30.297	3.16764	32.025	2.49383	34.291	2.92627	26.066	2.11641
96% Sat.	29.658	3.37324	30.816	2.20709	33.13	2.92678	25.052	2.13235
97% Sat.	29.507	3.43335	30.373	2.29075	32.926	2.97326	24.617	2.18032
98% Sat.	27.337	3.65683	25.903	2.80253	30.404	3.45576	21.57	2.51531
99% Sat.	27.12	3.68203	25.399	2.95823	30.143	3.47099	21.198	2.59388
100% Sat.	27.057	3.71933	25.245	2.8641	29.986	3.38536	21.044	2.57692

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	37.755	3.05344	35.9	3.49832	37.199	1.95449	37.103	3.00175	36.946	2.66312
92% Sat.	36.98	2.97054	35.289	3.3701	36.311	1.72274	36.182	3.01223	35.86	2.2873
93% Sat.	36.608	3.25396	34.975	3.38052	35.974	1.86731	36.032	3.13025	35.301	2.19755
94% Sat.	35.747	3.35283	33.823	3.52574	35.213	2.02249	35.358	3.42684	33.874	2.23685
95% Sat.	34.343	3.62176	32.498	3.84048	33.989	2.47884	34.667	3.72174	32.312	2.17334
96% Sat.	32.912	3.58113	31.649	3.73777	32.924	2.51497	33.134	3.66947	30.997	2.16176
97% Sat.	32.684	3.58972	31.472	3.74018	32.748	2.51755	32.905	3.68038	30.712	2.21773
98% Sat.	30.645	3.89841	29.706	4.02603	29.77	3.3034	30.864	4.00115	27.127	2.71333
99% Sat.	30.429	3.93173	29.281	4.26705	29.279	3.57127	30.617	4.02983	26.68	2.88846
100% Sat.	30.338	3.96907	29.126	4.20063	28.977	3.4389	30.54	4.07515	26.516	2.83455

Table H.18 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.959	0.25709	32.688	2.63528	18.309	1.12986	3.924	0.80305
92% Sat.	0.834	0.21413	31.727	2.18689	16.749	1.1968	3.428	0.73721
93% Sat.	0.815	0.20255	31.452	2.12501	16.476	1.19323	3.364	0.72165
94% Sat.	0.752	0.19615	30.333	2.20235	13.967	1.07861	2.943	0.65477
95% Sat.	0.331	0.13993	26.641	2.26955	9.368	1.1466	1.571	0.55009
96% Sat.	0.202	0.08418	25.111	1.99907	7.264	1.17024	0.968	0.35447
97% Sat.	0.154	0.04586	24.65	2.01217	6.271	1.16758	0.743	0.25961
98% Sat.	0.126	0.0437	19.361	2.02513	4.149	0.9572	0.614	0.24172
99% Sat.	0.125	0.04382	18.969	2.08522	4.104	0.93771	0.611	0.24183
100% Sat.	0.123	0.04327	18.961	2.08083	4.1	0.93711	0.61	0.24104

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	28.617	2.28116	32.06	3.22535	34.127	1.83318	33.493	2.74049	32.102	2.18118
92% Sat.	27.353	2.1951	31.49	3.05084	33.262	1.6259	32.64	2.72976	30.977	1.79964
93% Sat.	27.219	2.19372	31.443	3.04365	33.176	1.71652	32.626	2.72979	30.749	1.75538
94% Sat.	25.401	2.26962	30.617	3.31376	32.517	1.9185	32.068	3.00524	29.182	1.93639
95% Sat.	19.62	2.28442	28.491	3.34497	30.496	2.30776	30.747	3.21519	24.709	1.71786
96% Sat.	16.877	2.267	27.869	3.34586	29.382	2.23084	29.637	3.3234	22.989	1.72107
97% Sat.	15.357	2.25559	27.789	3.35368	29.237	2.27952	29.575	3.34191	22.466	1.7372
98% Sat.	10.46	1.74783	26.13	3.51494	26.108	2.9226	27.828	3.64107	17.76	1.70766
99% Sat.	10.078	1.74269	25.79	3.71957	25.775	3.11058	27.648	3.66091	17.505	1.74135
100% Sat.	10.071	1.74132	25.784	3.71796	25.761	3.10894	27.642	3.65991	17.501	1.73999

Table H.19 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	2.756	0.98341	68.326	0.74931	37.51	3.68097	10.165	2.95274
92% Sat.	2.411	0.9304	68.021	0.72906	34.448	3.75563	9.065	2.81227
93% Sat.	2.16	0.80217	67.726	0.69981	31.83	3.8224	8.029	2.48511
94% Sat.	1.999	0.72117	67.225	0.74157	28.289	3.68614	7.298	2.32751
95% Sat.	1.116	0.48144	62.554	1.178	20.547	3.38926	4.422	1.78915
96% Sat.	0.76	0.22914	61.049	1.39132	17.409	3.30462	3.423	1.47164
97% Sat.	0.55	0.17682	58.448	1.33911	13.065	2.77607	2.426	1.11635
98% Sat.	0.474	0.1417	56.438	1.61149	10.463	2.18822	2.049	0.88094
99% Sat.	0.437	0.14167	51.937	1.58064	9.684	2.22667	1.996	0.8816
100% Sat.	0.406	0.13628	51.744	1.63776	9.619	2.22603	1.946	0.8797

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	36.783	2.62096	63.781	1.00539	70.816	0.73336	30.741	2.03404	61.258	0.9136
92% Sat.	35.58	2.44711	63.694	0.99832	70.695	0.72566	29.86	2.02729	61.057	0.90455
93% Sat.	34.698	2.31484	63.644	0.99665	70.605	0.72048	28.726	1.76514	60.715	0.88736
94% Sat.	33.581	2.23906	63.489	0.94838	70.443	0.65245	28.243	1.84243	60.478	0.8493
95% Sat.	30.985	1.96487	62.165	0.85091	68.381	1.04515	26.046	1.83109	59.034	1.08851
96% Sat.	29.897	1.71527	61.654	0.86465	67.695	0.95086	25.449	1.82285	58.431	1.20346
97% Sat.	28.092	1.48754	61.23	1.17181	66.879	1.06538	24.418	1.55904	57.657	1.2285
98% Sat.	27.594	1.31228	60.924	1.06842	66.309	1.1728	24.039	1.63658	57.204	1.3337
99% Sat.	25.16	1.43853	60.251	1.41457	64.738	1.28993	23.94	1.64383	56.301	1.44806
100% Sat.	25.103	1.44789	60.135	1.419	64.549	1.27535	23.891	1.64893	56.178	1.44336

Table H.20 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	25.022	2.26871	30.084	1.36262	31.054	1.99683	25.081	1.77195
92% Sat.	25.015	2.26563	29.283	1.14154	30.353	1.85629	24.923	1.75952
93% Sat.	25.008	2.26122	29.067	1.24566	30.151	1.75253	24.721	1.8738
94% Sat.	24.522	2.68351	28.409	1.47864	29.676	1.93791	23.588	1.86532
95% Sat.	23.507	3.15269	26.435	1.74409	28	1.96589	22.194	1.7937
96% Sat.	23.114	3.50278	24.825	2.15556	27.128	2.48986	21.22	1.96951
97% Sat.	22.943	3.69905	24.339	2.39666	26.937	2.77148	20.873	2.12702
98% Sat.	21.07	4.19882	21.254	3.25716	25.278	2.89357	17.675	2.55301
99% Sat.	21.066	4.1993	21.08	3.33924	25.229	2.9272	17.564	2.60049
100% Sat.	21.006	4.1596	20.738	3.33223	25.148	2.86982	17.255	2.59997

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	29.153	2.18328	29.96	1.2423	31.645	1.45549	26.256	1.654	29.556	1.56448
92% Sat.	28.617	1.99222	29.742	1.09457	30.728	1.07743	25.893	1.27622	28.813	1.23346
93% Sat.	28.404	2.09635	29.568	1.13873	30.505	1.19805	25.749	1.27392	28.517	1.30155
94% Sat.	28.071	2.0216	29.192	1.09404	29.787	1.54633	24.811	1.45921	27.844	1.58592
95% Sat.	26.933	2.30925	28.198	0.88074	28.336	1.82842	23.414	1.65319	26.175	1.94522
96% Sat.	25.991	2.39208	28.042	0.8882	27.151	2.35448	22.758	1.85572	24.742	2.06727
97% Sat.	25.51	2.58242	27.821	0.9579	26.674	2.72233	22.512	2.07132	24.426	2.32232
98% Sat.	23.23	2.64292	26.23	0.88368	24.205	2.96911	19.355	2.12211	21.461	2.76014
99% Sat.	23.079	2.78515	26.202	0.9033	24.093	2.96202	19.258	2.12612	21.272	2.84422
100% Sat.	22.645	2.68629	26.144	0.87563	23.699	3.03817	18.862	2.22988	20.946	2.78539

Table H.21 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.846	0.69554	27.316	1.31858	15.649	2.6222	3.625	1.84166
92% Sat.	0.722	0.58506	26.578	1.31323	14.344	2.5697	3.153	1.70683
93% Sat.	0.722	0.58535	26.376	1.35782	14.295	2.56493	3.119	1.68099
94% Sat.	0.642	0.47671	25.539	1.51733	12.055	2.57044	2.572	1.45712
95% Sat.	0.383	0.36518	22.183	1.77665	8.324	2.27164	1.54	1.1394
96% Sat.	0.167	0.11465	20.254	2.07133	6.532	2.17078	0.993	0.87477
97% Sat.	0.131	0.10796	19.621	2.20016	5.712	2.02996	0.81	0.75427
98% Sat.	0.099	0.07116	15.91	2.95324	3.573	1.37389	0.555	0.48499
99% Sat.	0.098	0.07084	15.843	2.93673	3.547	1.37547	0.553	0.4839
100% Sat.	0.097	0.07066	15.834	2.93622	3.543	1.37445	0.552	0.4842

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	4.608	2.11436	26.225	1.1713	29.544	1.44991	18.139	1.91836	27.327	1.41967
92% Sat.	3.986	1.95019	26.08	0.98616	28.701	1.17692	17.939	1.93419	26.683	1.16821
93% Sat.	3.951	1.95513	25.987	0.99385	28.613	1.29832	17.912	1.91858	26.443	1.24372
94% Sat.	3.317	1.68742	25.535	0.93128	27.922	1.61664	17.628	1.97957	25.749	1.52796
95% Sat.	2.107	1.32372	24.144	0.92659	25.764	1.85104	16.901	1.92751	23.771	1.92436
96% Sat.	1.331	0.94297	23.705	0.99246	24.341	2.21501	16.544	1.95238	22.248	1.99463
97% Sat.	1.09	0.82642	23.411	1.05581	23.821	2.56066	16.413	1.92257	21.936	2.20366
98% Sat.	0.757	0.56337	21.75	0.99067	21.267	2.93494	16.129	1.9699	18.989	2.64283
99% Sat.	0.754	0.56259	21.727	1.00727	21.196	2.9352	16.122	1.96963	18.916	2.63525
100% Sat.	0.752	0.56274	21.724	1.00722	21.184	2.9368	16.116	1.96855	18.904	2.63053

Table H.22 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.856	0.31656	45.87	1.53697	15.184	2.7571	2.93	1.08651
92% Sat.	0.774	0.3094	45.533	1.63015	13.629	2.29088	2.638	1.03472
93% Sat.	0.734	0.31343	45.016	1.76136	13.038	2.15241	2.592	1.03524
94% Sat.	0.693	0.3128	44.484	1.73695	12.549	2.08736	2.559	1.02831
95% Sat.	0.379	0.10463	40.015	2.20101	7.009	1.7294	1.056	0.4364
96% Sat.	0.304	0.05976	38.496	2.29651	5.972	1.6566	0.895	0.34416
97% Sat.	0.244	0.06435	37.216	2.54787	4.504	1.18915	0.759	0.30515
98% Sat.	0.208	0.06596	34.681	2.65663	4.28	1.13097	0.704	0.30341
99% Sat.	0.159	0.06142	32.981	2.4338	4.245	1.13412	0.655	0.30011
100% Sat.	0.132	0.06137	32.917	2.43005	4.199	1.13432	0.606	0.30081

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	42.089	2.47936	3.581	1.35747	3.304	1.22011	23.291	2.28467	45.875	1.51765
92% Sat.	42.013	2.49045	3.221	1.29002	2.977	1.17884	21.888	1.93714	45.526	1.59791
93% Sat.	41.873	2.47553	3.161	1.27509	2.921	1.16856	21.294	1.8109	45.047	1.72043
94% Sat.	41.708	2.51484	3.126	1.2711	2.891	1.1665	20.794	1.75843	44.51	1.71752
95% Sat.	40.92	2.53244	1.596	0.66462	1.185	0.48051	17.926	2.14058	40.044	2.18829
96% Sat.	40.466	2.39199	1.439	0.63347	1.002	0.42719	17.264	2.09788	38.501	2.33822
97% Sat.	40.237	2.49003	1.29	0.6224	0.844	0.40159	15.847	1.7509	37.227	2.56962
98% Sat.	39.502	2.42278	1.241	0.62461	0.802	0.39499	15.61	1.71811	34.716	2.7089
99% Sat.	38.934	2.21335	1.194	0.62207	0.761	0.39248	15.588	1.7178	33.009	2.46849
100% Sat.	38.906	2.21668	1.138	0.62086	0.703	0.39277	15.545	1.72162	32.951	2.4639

Table H.23 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	76.211	2.38182	76.148	1.57511	80.219	1.22478	75.475	1.52662
92% Sat.	76.168	2.37571	75.829	1.61075	80.046	1.2302	75.174	1.56729
93% Sat.	76.114	2.37249	75.149	1.64523	79.825	1.21502	74.536	1.58329
94% Sat.	76.091	2.3714	74.893	1.639	79.607	1.25909	74.235	1.59546
95% Sat.	75.989	2.36545	73.889	1.83224	79.147	1.34077	73.127	1.72671
96% Sat.	75.901	2.37462	73.21	1.80137	78.881	1.28686	72.346	1.65493
97% Sat.	75.814	2.38654	72.085	1.70544	78.257	1.25679	70.935	1.62554
98% Sat.	75.532	2.33798	70.984	1.58715	77.491	1.2208	69.636	1.38824
99% Sat.	75.132	2.48354	68.668	1.68574	76.089	1.25779	66.69	1.39519
100% Sat.	74.702	2.42129	67.253	1.73036	74.832	1.28349	65.185	1.50276

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	80.984	1.27411	76.989	1.77462	76.429	1.47554	78.566	1.30627	76.167	1.72015
92% Sat.	80.864	1.28361	76.824	1.80798	76.143	1.49099	78.355	1.32455	75.868	1.78005
93% Sat.	80.686	1.29548	76.378	1.83093	75.613	1.50037	77.97	1.33096	75.171	1.74767
94% Sat.	80.509	1.34231	76.155	1.81918	75.298	1.51591	77.661	1.37669	74.893	1.75043
95% Sat.	80.098	1.39444	75.348	1.91539	74.296	1.57931	77.035	1.41188	73.941	1.86972
96% Sat.	79.853	1.33106	74.746	1.81513	73.692	1.52262	76.658	1.36155	73.294	1.83284
97% Sat.	79.39	1.3102	73.858	1.78104	72.446	1.57668	75.714	1.47314	72.056	1.76478
98% Sat.	78.882	1.20659	72.948	1.60544	71.343	1.40316	74.772	1.40781	70.948	1.57087
99% Sat.	77.516	1.28189	70.613	1.59374	68.711	1.50456	72.681	1.39137	68.737	1.68268
100% Sat.	76.793	1.26049	69.149	1.72504	67.237	1.59384	71.194	1.45339	67.337	1.77802

Table H.24 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.427	0.31831	45.228	1.52695	14.67	2.74743	2.482	1.07789
92% Sat.	0.385	0.30785	44.882	1.62421	13.127	2.2804	2.215	1.03135
93% Sat.	0.378	0.30768	44.363	1.75051	12.551	2.14045	2.19	1.03059
94% Sat.	0.368	0.30656	43.832	1.72748	12.075	2.07651	2.18	1.02299
95% Sat.	0.113	0.11253	39.363	2.18986	6.554	1.72024	0.712	0.42889
96% Sat.	0.07	0.05455	37.842	2.28484	5.516	1.64634	0.56	0.33249
97% Sat.	0.064	0.05165	36.575	2.5389	4.079	1.18012	0.479	0.30127
98% Sat.	0.056	0.05382	34.058	2.63875	3.88	1.12514	0.469	0.30452
99% Sat.	0.052	0.05031	32.374	2.42209	3.873	1.12558	0.462	0.30052
100% Sat.	0.051	0.05064	32.355	2.42914	3.864	1.12937	0.457	0.2994

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	41.492	2.46561	3.099	1.3527	2.846	1.21629	22.78	2.28111	45.245	1.50197
92% Sat.	41.408	2.47853	2.765	1.28593	2.547	1.17176	21.383	1.92961	44.882	1.58328
93% Sat.	41.269	2.46295	2.723	1.26639	2.51	1.16172	20.797	1.80237	44.398	1.70237
94% Sat.	41.103	2.4989	2.713	1.26327	2.5	1.15835	20.308	1.74729	43.864	1.69684
95% Sat.	40.32	2.51542	1.223	0.65988	0.832	0.48095	17.462	2.13192	39.398	2.1656
96% Sat.	39.856	2.37247	1.075	0.63278	0.663	0.41699	16.798	2.08787	37.849	2.31584
97% Sat.	39.634	2.46657	0.987	0.62367	0.565	0.38834	15.412	1.74397	36.592	2.54789
98% Sat.	38.896	2.40362	0.977	0.62577	0.556	0.38915	15.188	1.70983	34.097	2.68449
99% Sat.	38.333	2.1949	0.974	0.62454	0.548	0.38691	15.182	1.70935	32.41	2.45331
100% Sat.	38.322	2.19471	0.971	0.62499	0.54	0.38702	15.169	1.71287	32.395	2.45721

Table H.25 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.896	0.30752	46.004	1.96025	15.205	1.73152	2.891	0.89376
92% Sat.	0.815	0.26069	45.437	2.09381	13.684	1.52367	2.719	0.79211
93% Sat.	0.779	0.25976	44.838	2.13983	13.199	1.42502	2.649	0.82068
94% Sat.	0.739	0.25483	44.197	1.9915	12.801	1.42418	2.616	0.81845
95% Sat.	0.352	0.08546	39.857	2.02183	6.934	1.13224	1.088	0.3683
96% Sat.	0.292	0.05881	38.443	1.6898	5.558	0.9867	0.857	0.29166
97% Sat.	0.242	0.04905	36.607	1.95582	4.602	0.8277	0.75	0.22546
98% Sat.	0.21	0.05115	33.929	2.02912	4.423	0.86685	0.704	0.2232
99% Sat.	0.164	0.05954	32.432	1.49075	4.392	0.87027	0.652	0.2323
100% Sat.	0.128	0.05739	32.382	1.48196	4.342	0.87078	0.594	0.24499

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	15.441	1.57063	21.704	2.19583	42.681	2.01858	30.439	1.9939	5.139	1.08079
92% Sat.	13.928	1.40934	20.343	2.11012	42.475	2.08655	27.217	1.98145	4.518	0.86522
93% Sat.	13.452	1.31846	19.841	2.04223	42.283	2.05042	25.81	1.88473	4.306	0.87677
94% Sat.	13.038	1.31388	19.38	2.07022	42.099	1.9916	23.614	1.72826	4.16	0.83471
95% Sat.	7.214	1.06437	14.5	2.12947	40.991	2.24698	15.297	1.49373	1.853	0.40461
96% Sat.	5.845	1.00409	13.286	1.91177	40.679	2.22592	12.202	1.25192	1.32	0.3893
97% Sat.	4.885	0.87141	12.178	1.74286	40.507	2.28496	9.599	1.01791	1.095	0.30195
98% Sat.	4.699	0.90863	11.803	1.66438	39.709	2.01455	8.413	0.96508	1.008	0.30296
99% Sat.	4.639	0.91757	11.656	1.5642	39.134	1.92553	8.047	0.9213	0.963	0.3063
100% Sat.	4.585	0.92339	11.615	1.56464	39.112	1.92306	7.995	0.92263	0.901	0.30501

Table H.26 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	65.694	2.80535	65.164	1.99218	69.234	1.81674	63.722	2.00981
92% Sat.	65.523	2.78444	64.601	2.01001	68.858	1.87431	63.136	2.0569
93% Sat.	65.386	2.74166	63.933	2.01821	68.508	1.95115	62.506	2.0536
94% Sat.	65.161	2.79046	63.521	2.02543	68.242	1.9941	62.042	2.06849
95% Sat.	64.739	2.81807	62.185	2.06591	67.358	2.16705	60.607	2.15875
96% Sat.	64.454	2.73528	61.422	2.05951	66.909	2.13145	59.617	2.26388
97% Sat.	64.185	2.74347	60.047	2.06084	66.176	2.24569	58.148	2.21202
98% Sat.	63.52	2.84191	58.489	2.19072	65.088	2.38466	56.243	2.21819
99% Sat.	62.97	2.82341	55.978	2.11915	63.53	2.6209	53.142	2.04513
100% Sat.	62.513	2.81816	54.318	2.13999	62.27	2.56319	51.538	2.05438

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	66.911	1.93536	67.409	2.81859	0.04	1.89539	67.986	2.67148	64.545	2.02929
92% Sat.	66.36	2.04804	67.221	2.87457	0.04	2.00833	67.751	2.69209	64.068	2.07596
93% Sat.	65.752	2.17919	67.106	2.88549	0.04	2.13977	67.608	2.6815	63.419	2.07777
94% Sat.	65.326	2.19741	66.904	2.92983	0.04	2.1582	67.431	2.71732	63.002	2.0369
95% Sat.	64.175	2.28463	66.268	3.0067	0.039	2.24604	66.815	2.85352	61.739	2.1314
96% Sat.	63.489	2.34564	65.986	2.95838	0.038	2.30744	66.489	2.78948	60.954	2.08622
97% Sat.	62.168	2.31252	65.576	3.01964	0.038	2.27501	66.156	2.86181	59.44	1.96984
98% Sat.	60.726	2.35498	64.925	3.05967	0.037	2.31826	65.577	2.86575	57.98	2.05809
99% Sat.	58.521	2.11966	63.25	3.13442	0.036	2.08435	64.889	2.85636	55.825	1.96518
100% Sat.	56.785	2.1525	61.935	3.16828	0.035	2.11811	64.343	2.95896	54.231	1.95061

Table H.27 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.418	0.2869	43.586	1.81906	14.348	1.73537	2.29	0.84941
92% Sat.	0.385	0.24663	43.047	1.91255	12.878	1.51818	2.169	0.75393
93% Sat.	0.382	0.2405	42.5	1.95217	12.429	1.41241	2.132	0.77736
94% Sat.	0.382	0.24083	41.954	1.83122	12.061	1.41656	2.129	0.77687
95% Sat.	0.072	0.06643	37.787	1.86248	6.319	1.11969	0.698	0.34086
96% Sat.	0.041	0.03058	36.456	1.55981	4.951	0.96359	0.484	0.26313
97% Sat.	0.039	0.0286	34.782	1.86162	4.048	0.81354	0.428	0.21128
98% Sat.	0.037	0.02826	32.234	1.92304	3.913	0.84489	0.427	0.21178
99% Sat.	0.035	0.02909	30.845	1.42602	3.911	0.84541	0.426	0.21185
100% Sat.	0.034	0.02926	30.843	1.42539	3.909	0.8447	0.425	0.21174

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	14.331	1.57183	19.044	2.14521	41.634	1.98937	27.761	1.84018	4.311	1.02336
92% Sat.	12.859	1.41943	17.679	2.05292	41.426	2.06716	24.529	1.83803	3.755	0.82636
93% Sat.	12.413	1.32221	17.237	1.98094	41.226	2.03461	23.175	1.72361	3.579	0.84236
94% Sat.	12.018	1.32653	16.82	1.99515	41.014	1.99611	20.998	1.5001	3.472	0.81447
95% Sat.	6.303	1.06701	11.96	2.04554	39.951	2.24547	12.79	1.2599	1.288	0.37499
96% Sat.	4.968	0.99544	10.763	1.77339	39.608	2.24976	9.733	1.06941	0.774	0.35082
97% Sat.	4.062	0.87982	9.702	1.58274	39.389	2.30562	7.188	0.83934	0.622	0.27205
98% Sat.	3.915	0.91045	9.478	1.53736	38.529	2.09385	6.094	0.86541	0.601	0.26297
99% Sat.	3.912	0.91136	9.471	1.53242	37.969	2.01362	5.859	0.87067	0.6	0.2632
100% Sat.	3.91	0.91112	9.47	1.53189	37.966	2.01304	5.858	0.87075	0.597	0.26234

Table H.28 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.961	0.35339	47.194	1.74399	16.731	3.11394	3.377	1.48752
92% Sat.	0.879	0.35982	46.75	1.79049	15.129	2.9357	3.068	1.32092
93% Sat.	0.842	0.36127	46.224	1.60577	14.647	2.91617	3.006	1.32732
94% Sat.	0.805	0.36182	45.692	1.60807	13.96	2.6979	2.976	1.31822
95% Sat.	0.396	0.15665	41.677	1.81752	7.774	2.38605	1.103	0.34196
96% Sat.	0.317	0.07642	39.602	1.81156	6.617	2.24391	0.928	0.34023
97% Sat.	0.264	0.08245	37.75	2.26264	5.388	1.86945	0.808	0.2605
98% Sat.	0.22	0.08493	35.424	2.09277	4.936	1.56221	0.764	0.2617
99% Sat.	0.179	0.09284	33.591	2.03304	4.906	1.56293	0.73	0.26295
100% Sat.	0.142	0.08609	33.542	2.02672	4.856	1.56232	0.68	0.26397

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	41.784	1.17778	39.211	1.71798	33.88	2.64268	46.583	1.62108	46.581	1.85171
92% Sat.	41.689	1.08612	39.088	1.70992	32.293	2.59748	46.389	1.57354	46.471	1.80748
93% Sat.	41.603	1.06871	38.941	1.67896	31.672	2.51164	45.933	1.44878	46.276	1.77008
94% Sat.	41.433	1.11938	38.81	1.67099	30.812	2.49203	45.457	1.29861	45.975	1.76528
95% Sat.	39.863	1.40303	37.296	1.60446	24.735	2.34631	42.869	1.25845	44.85	1.65571
96% Sat.	39.506	1.26498	36.918	1.53408	22.958	2.18538	41.481	0.99836	44.227	1.51623
97% Sat.	39.12	1.26603	36.399	1.55555	20.862	1.89853	40.669	1.28115	43.822	1.52674
98% Sat.	38.872	1.2184	35.889	1.55019	19.598	1.76421	39.422	1.20425	43.441	1.53416
99% Sat.	38.557	1.21881	35.322	1.47515	18.579	1.66075	37.854	0.90379	42.366	1.44403
100% Sat.	38.532	1.20934	35.296	1.46938	18.528	1.65825	37.817	0.8964	42.338	1.43662

Table H.29 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	53.538	2.54348	54.202	2.84387	58.193	3.09765	52.195	2.74303
92% Sat.	53.26	2.53452	53.649	2.81001	57.73	3.12751	51.638	2.55048
93% Sat.	52.9	2.46302	52.755	3.09644	57.098	3.08446	50.789	2.68465
94% Sat.	52.528	2.49458	52.228	3.1228	56.638	3.18816	50.289	2.62902
95% Sat.	51.897	2.32385	50.741	2.78669	55.412	3.09984	48.737	2.29426
96% Sat.	51.469	2.3202	49.645	2.91523	54.849	3.20569	47.634	2.26939
97% Sat.	51.103	2.13833	48.375	2.82598	53.795	3.05913	46.177	2.33266
98% Sat.	50.318	2.22977	46.921	3.0826	52.526	3.14198	44.699	2.55159
99% Sat.	49.608	1.97453	44.571	2.74847	51.06	2.84646	41.782	2.14943
100% Sat.	49.197	1.88875	43.977	2.70629	49.933	2.82236	41.159	2.07928

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	54.872	2.93606	56.52	3.59541	54.881	2.95879	57.702	3.09955	58.204	2.94421
92% Sat.	54.331	2.9015	56.051	3.70639	54.365	2.90504	57.302	3.12081	57.733	3.01864
93% Sat.	53.703	3.03587	55.542	3.81897	53.665	3.103	56.989	3.06888	57.488	3.04137
94% Sat.	53.24	3.12182	55.249	3.92586	53.198	3.21302	56.6	3.17196	57.161	3.15043
95% Sat.	51.9	2.74045	54.187	3.55416	51.747	2.88582	55.713	2.93993	56.233	3.06597
96% Sat.	50.863	2.70724	53.354	3.61681	50.823	3.00869	55.214	2.84577	55.623	3.02518
97% Sat.	49.636	2.49638	52.621	3.40671	49.818	2.76935	54.676	2.78241	55.03	2.85015
98% Sat.	48.176	2.78113	51.234	3.76857	48.533	3.15789	53.603	2.78208	54.045	2.83104
99% Sat.	45.85	2.54642	49.362	3.84602	46.403	2.7581	52.734	2.61666	53.19	2.6551
100% Sat.	44.996	2.41898	48.176	3.68263	45.577	2.64959	52.169	2.53695	52.569	2.64569

Table H.30 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.481	0.33542	41.08	1.76666	14.94	2.87995	2.654	1.414
92% Sat.	0.448	0.34302	40.669	1.7071	13.524	2.71715	2.408	1.2607
93% Sat.	0.44	0.34906	40.19	1.53023	13.122	2.71762	2.382	1.26874
94% Sat.	0.434	0.34379	39.72	1.55323	12.504	2.48953	2.369	1.25381
95% Sat.	0.101	0.17602	36.1	1.60233	6.781	2.20797	0.674	0.30998
96% Sat.	0.053	0.07964	34.243	1.66009	5.675	2.07414	0.535	0.31259
97% Sat.	0.05	0.08014	32.715	2.03342	4.606	1.73007	0.48	0.24925
98% Sat.	0.048	0.07921	30.732	1.88261	4.223	1.4349	0.478	0.24852
99% Sat.	0.047	0.0794	29.217	1.79015	4.22	1.43611	0.477	0.24797
100% Sat.	0.047	0.07948	29.214	1.78955	4.215	1.43523	0.475	0.24754

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	35.794	1.2432	38.03	1.77415	27.679	2.43312	40.116	1.85282	40.659	1.9525
92% Sat.	35.546	1.14109	37.935	1.75915	25.995	2.31962	39.778	1.84983	40.393	1.95168
93% Sat.	35.516	1.13028	37.798	1.72895	25.378	2.23561	39.355	1.70214	40.215	1.93077
94% Sat.	35.316	1.14335	37.655	1.73827	24.494	2.27144	38.854	1.67011	39.844	1.93287
95% Sat.	33.623	1.34485	36.296	1.62593	18.289	2.06886	36.152	1.54543	38.633	1.73237
96% Sat.	33.244	1.26348	35.93	1.54496	16.52	1.96049	34.793	1.32398	38.053	1.70529
97% Sat.	32.79	1.23825	35.319	1.54117	14.414	1.60434	33.944	1.41906	37.547	1.57456
98% Sat.	32.461	1.10102	34.558	1.4832	13.086	1.46653	32.577	1.27181	37.077	1.50751
99% Sat.	32.306	1.11378	33.973	1.41611	12.213	1.46149	31.301	1.04058	36.271	1.56002
100% Sat.	32.296	1.10606	33.967	1.41102	12.208	1.46036	31.298	1.0364	36.269	1.55851

Table H.31 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP / Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.85	0.25514	46.661	1.65113	16.764	1.70678	3.446	1.00751
92% Sat.	0.791	0.24921	46.337	1.73053	14.971	1.41961	3.153	1.00526
93% Sat.	0.745	0.23589	45.711	1.6801	14.678	1.34126	3.067	0.96326
94% Sat.	0.712	0.23287	45.043	1.71453	14.34	1.38032	3.032	0.95967
95% Sat.	0.365	0.07549	40.997	1.87051	8.145	1.32345	1.169	0.43385
96% Sat.	0.318	0.07972	39.035	2.12368	6.749	1.1689	0.948	0.3378
97% Sat.	0.259	0.07249	37.024	2.08318	5.563	0.87558	0.808	0.31807
98% Sat.	0.216	0.07112	34.648	1.92072	5.398	0.86314	0.763	0.32028
99% Sat.	0.176	0.07299	32.845	1.4217	5.365	0.86033	0.729	0.32042
100% Sat.	0.142	0.0668	32.799	1.40527	5.317	0.85454	0.687	0.30365

Policies / DP / Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	46.606	1.604	44.295	1.66328	45.77	2.22653	33.837	1.46073	30.587	2.19655
92% Sat.	46.282	1.68879	43.62	1.72108	45.635	2.24312	32.219	1.55442	29.717	2.22061
93% Sat.	45.678	1.61792	42.865	1.56693	45.317	2.23899	31.473	1.5061	29.21	2.04694
94% Sat.	45.048	1.64855	41.477	1.55447	45.151	2.25111	30.476	1.47227	28.518	2.04286
95% Sat.	41.047	1.82018	39.109	1.51703	43.531	2.16723	25.008	1.35175	23.717	2.1558
96% Sat.	39.151	2.01129	37.444	1.53521	42.653	2.27339	22.926	1.2838	22.351	2.27403
97% Sat.	37.179	1.98613	35.102	1.66111	42.079	2.19841	20.854	1.33297	20.802	2.17469
98% Sat.	34.798	1.85062	33.063	1.60674	41.215	2.08533	19.638	1.25369	19.299	2.10385
99% Sat.	32.99	1.39576	32.45	1.4789	39.85	1.94611	18.882	0.89657	18.973	2.09693
100% Sat.	32.942	1.3792	32.417	1.47556	39.82	1.93512	18.837	0.90167	18.944	2.09908

Table H.32 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	36.033	2.91661	38.665	2.91698	41.611	4.1341	35.669	2.45518
92% Sat.	35.771	2.87537	38.213	2.84537	41.164	4.21793	35.154	2.34165
93% Sat.	35.327	2.85098	36.887	2.83433	40.299	4.19179	34.152	2.22465
94% Sat.	34.674	2.84298	36.285	2.70268	39.791	3.82412	33.579	2.05732
95% Sat.	33.56	2.34658	34.393	2.20703	37.861	3.27904	31.648	1.55341
96% Sat.	32.712	2.58855	33.201	2.18033	36.813	3.16379	30.497	1.91434
97% Sat.	32.04	2.28201	32.002	1.64714	35.542	2.66576	29.316	1.49418
98% Sat.	31.494	2.0061	31.072	1.6181	34.45	2.77158	28.097	1.18918
99% Sat.	30.986	1.84072	29.72	1.10927	33.369	2.49172	26.877	1.13427
100% Sat.	30.887	1.81961	29.589	0.98568	32.978	2.25903	26.811	1.14679

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	38.643	2.78659	41.559	3.67278	40.861	3.62031	40.782	3.53431	37.37	2.27186
92% Sat.	38.194	2.70829	41.211	3.59679	40.396	3.50952	40.423	3.67442	36.852	2.27571
93% Sat.	36.763	2.5023	40.627	3.61838	39.274	3.28209	40.014	3.6758	35.925	2.28363
94% Sat.	36.154	2.45628	40.017	3.40362	38.378	2.80103	39.302	3.709	35.29	2.19013
95% Sat.	34.335	2.04618	38.446	3.12478	36.426	2.48961	37.709	3.49199	33.664	1.77364
96% Sat.	33.245	2.22066	37.281	3.1643	35.285	2.57365	36.681	3.61882	32.623	1.96823
97% Sat.	32.048	1.71201	35.924	2.92039	33.854	2.12801	35.7	3.4903	31.441	1.46973
98% Sat.	31.143	1.62796	34.902	2.76698	32.745	1.94472	34.777	3.31106	30.584	1.28781
99% Sat.	29.748	1.06384	33.876	2.27012	31.206	1.6468	33.777	2.85597	29.31	0.82917
100% Sat.	29.657	0.98297	33.549	2.1831	31.054	1.53421	33.528	2.91464	29.198	0.79567

Table H.33 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.339	0.21676	31.993	1.55409	12.564	1.07527	2.288	0.78251
92% Sat.	0.324	0.2169	31.756	1.55133	11.38	0.92145	2.108	0.82297
93% Sat.	0.317	0.20142	31.19	1.488	11.206	0.91018	2.056	0.79643
94% Sat.	0.316	0.2013	30.557	1.38865	10.96	0.90105	2.054	0.79685
95% Sat.	0.062	0.04783	27.055	1.34521	5.847	0.95359	0.557	0.32547
96% Sat.	0.045	0.05152	25.576	1.356	4.701	0.80985	0.448	0.24742
97% Sat.	0.045	0.05035	24.132	1.52671	3.866	0.66075	0.372	0.25603
98% Sat.	0.042	0.05032	22.607	1.50592	3.787	0.65143	0.37	0.2561
99% Sat.	0.042	0.05059	21.713	1.13488	3.786	0.65133	0.369	0.25597
100% Sat.	0.042	0.05059	21.711	1.13438	3.784	0.65178	0.369	0.25627

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	31.986	1.39227	30.393	0.99816	33.133	1.67881	33.837	1.46073	20.165	1.89407
92% Sat.	31.741	1.43332	29.681	1.04887	32.995	1.66239	32.219	1.55442	19.548	1.88312
93% Sat.	31.158	1.33426	28.909	0.89049	32.495	1.71827	31.473	1.5061	19.209	1.88061
94% Sat.	30.517	1.26665	27.224	0.67525	31.889	1.355	30.476	1.47227	18.676	1.77459
95% Sat.	27.076	1.19234	24.593	0.63029	30.173	1.46513	25.008	1.35175	14.777	1.81694
96% Sat.	25.71	1.29481	22.801	0.77538	29.199	1.55809	22.926	1.2838	13.789	2.05014
97% Sat.	24.3	1.38016	20.23	1.14907	28.564	1.45327	20.854	1.33297	12.657	1.96765
98% Sat.	22.793	1.37639	18.133	0.8867	27.635	1.19787	19.638	1.25369	11.901	1.821
99% Sat.	21.861	1.04959	18.031	0.8415	26.921	1.02805	18.882	0.89657	11.793	1.73846
100% Sat.	21.859	1.04896	18.03	0.84153	26.919	1.0289	18.837	0.90167	11.791	1.73829

Table H.34 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.796	0.22928	46.61	1.84559	16.404	2.09127	2.857	0.69764
92% Sat.	0.726	0.23268	46.135	1.73251	14.619	1.91753	2.62	0.64162
93% Sat.	0.681	0.2327	45.525	1.58584	14.131	1.87881	2.506	0.68153
94% Sat.	0.641	0.22852	44.894	1.42304	13.531	1.83962	2.462	0.67614
95% Sat.	0.347	0.04394	40.532	1.76946	7.626	1.58743	1.005	0.27439
96% Sat.	0.295	0.0322	38.602	1.69258	6.336	1.34457	0.766	0.17182
97% Sat.	0.243	0.03289	36.801	1.55635	4.901	1.07047	0.688	0.15634
98% Sat.	0.195	0.03586	34.454	1.09906	4.529	1.00961	0.634	0.15739
99% Sat.	0.147	0.04632	32.752	1.24157	4.49	1.00822	0.592	0.15961
100% Sat.	0.111	0.04673	32.693	1.23971	4.438	1.01384	0.546	0.16693

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	36.939	2.19108	48.06	2.11101	32.331	1.94609	3.325	0.74293	46.624	1.87254
92% Sat.	35.538	2.14133	47.897	2.12617	31.946	1.90603	3.05	0.66253	46.146	1.75004
93% Sat.	34.947	1.99804	47.656	2.10423	31.854	1.84747	2.923	0.71592	45.562	1.66162
94% Sat.	34.099	1.78317	47.437	1.98232	31.555	1.73587	2.887	0.70877	44.909	1.52029
95% Sat.	28.124	1.85266	45.597	2.31951	29.547	1.7455	1.205	0.27455	40.467	1.81669
96% Sat.	26.147	1.6541	44.795	2.07544	29.081	1.72253	0.908	0.22565	38.587	1.75802
97% Sat.	23.894	1.40429	43.978	1.73192	28.426	1.57929	0.817	0.18925	36.756	1.59836
98% Sat.	21.877	1.53612	42.956	1.55994	28.116	1.56801	0.778	0.17984	34.452	1.21383
99% Sat.	20.665	1.45696	41.9	1.32676	27.578	1.3766	0.746	0.18017	32.751	1.27001
100% Sat.	20.629	1.45434	41.866	1.32915	27.562	1.37387	0.7	0.18649	32.692	1.26515

Table H.35 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	23.998	2.44491	27.152	3.29357	27.841	3.29734	23.905	3.0187
92% Sat.	23.638	2.35724	26.3	3.31044	27.267	3.28502	23.202	2.8368
93% Sat.	23.05	2.14726	25.357	3.34268	26.567	3.31247	22.593	2.55462
94% Sat.	22.157	2.0963	25.002	3.22685	26.253	3.35644	22.291	2.38232
95% Sat.	21.211	1.84738	23.461	3.56027	24.619	3.13564	20.507	2.20077
96% Sat.	20.451	1.80662	22.685	3.3574	23.801	2.95133	19.829	1.96452
97% Sat.	19.83	1.62619	21.422	2.53377	22.715	2.45927	18.475	1.49073
98% Sat.	19.538	1.48751	20.726	2.19355	22.01	2.20004	17.728	1.45638
99% Sat.	19.127	1.15343	19.941	1.61912	21.296	1.86052	17.457	1.29057
100% Sat.	19.109	1.13191	19.931	1.62307	21.242	1.88767	17.445	1.29397

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	26.457	3.64287	28.49	3.34471	26.699	3.0266	25.329	3.52668	27.221	2.98117
92% Sat.	25.573	3.4987	27.742	3.32168	26.149	3.03513	24.726	3.42102	26.394	2.96661
93% Sat.	24.648	3.30097	26.719	3.44465	25.608	2.85998	24.181	3.38009	25.44	3.11793
94% Sat.	24.264	2.99011	26.016	3.29224	24.874	3.30824	23.908	3.26784	24.873	3.07768
95% Sat.	22.822	2.87555	24.516	3.44084	23.159	3.06522	21.928	3.28836	23.366	3.42264
96% Sat.	21.989	2.54888	23.652	3.27968	22.429	2.91415	21.057	2.92282	22.691	3.25037
97% Sat.	21.159	2.08294	22.283	2.58031	21.637	2.67312	19.676	2.12233	21.502	2.48826
98% Sat.	20.435	1.80445	21.579	2.27812	21.237	2.6703	18.78	1.84694	20.827	2.19536
99% Sat.	19.928	1.52944	20.889	1.66168	20.62	2.24797	18.382	1.65299	20.031	1.61189
100% Sat.	19.883	1.54198	20.868	1.67592	20.597	2.21376	18.358	1.63205	20.024	1.6153

Table H.36 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.168	0.12946	23.534	1.95994	9.667	1.72419	1.428	0.59421
92% Sat.	0.158	0.13113	22.89	1.9463	8.671	1.5691	1.34	0.57576
93% Sat.	0.158	0.13144	22.417	1.86313	8.5	1.55017	1.303	0.59836
94% Sat.	0.156	0.13118	22.025	1.84297	8.213	1.45903	1.292	0.59738
95% Sat.	0.041	0.03837	18.488	1.60652	4.135	1.24252	0.398	0.21233
96% Sat.	0.029	0.02963	17.402	1.51415	3.321	1.00332	0.273	0.1628
97% Sat.	0.028	0.02984	16.168	1.46622	2.629	0.82636	0.258	0.15598
98% Sat.	0.026	0.03049	15.008	1.52534	2.486	0.77421	0.253	0.15606
99% Sat.	0.025	0.03053	14.66	1.29877	2.485	0.77429	0.253	0.15612
100% Sat.	0.025	0.03063	14.659	1.29925	2.484	0.77456	0.252	0.15617

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	21.756	2.14461	24.92	2.08677	23.043	1.92492	1.721	0.73949	23.677	1.83639
92% Sat.	20.846	1.96896	24.498	2.1113	22.671	1.90004	1.594	0.67971	23.03	1.83528
93% Sat.	20.338	1.74274	24.158	2.23146	22.528	1.94301	1.546	0.70705	22.524	1.72359
94% Sat.	19.84	1.60518	23.582	2.08304	21.839	2.39761	1.535	0.70217	21.928	1.7418
95% Sat.	15.389	1.37758	21.451	1.91368	19.949	2.1915	0.497	0.25984	18.442	1.55707
96% Sat.	13.895	1.19147	20.514	1.80847	19.326	2.06995	0.328	0.18953	17.404	1.46002
97% Sat.	12.589	1.08871	19.529	1.61717	18.722	2.02737	0.301	0.16661	16.206	1.3926
98% Sat.	11.287	1.07676	18.695	1.51462	18.359	2.04844	0.296	0.16459	15.087	1.47537
99% Sat.	11.131	1.04145	18.395	1.19918	18.276	1.9443	0.295	0.16471	14.716	1.33179
100% Sat.	11.13	1.04244	18.394	1.19916	18.275	1.94418	0.295	0.16507	14.714	1.33224

Table H.37 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.832	0.1681	45.992	1.5358	16.272	2.59055	2.948	0.75182
92% Sat.	0.718	0.14698	45.526	1.59942	14.765	2.35385	2.656	0.73325
93% Sat.	0.678	0.15201	44.999	1.53019	14.38	2.30543	2.592	0.68692
94% Sat.	0.639	0.15876	44.424	1.49183	13.771	2.13703	2.528	0.67364
95% Sat.	0.362	0.06282	40.155	2.01809	7.833	1.996	1.152	0.43307
96% Sat.	0.317	0.0596	38.277	1.80984	6.453	1.5439	0.971	0.33381
97% Sat.	0.242	0.04064	36.423	1.96276	5.124	1.14663	0.775	0.2302
98% Sat.	0.197	0.03557	33.751	2.20233	4.731	1.02534	0.699	0.20454
99% Sat.	0.152	0.03956	32.302	1.94904	4.68	1.02876	0.65	0.20668
100% Sat.	0.121	0.03784	32.246	1.94989	4.633	1.02544	0.601	0.20883

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	23.639	1.20996	39.773	1.79024	47.345	1.87885	16.03	1.44308	34.554	2.15139
92% Sat.	23.205	1.31718	39.655	1.82127	47.137	1.91297	15.753	1.35704	34.47	2.16183
93% Sat.	23.028	1.32306	39.494	1.74188	46.851	1.87337	15.7	1.35947	34.282	2.1003
94% Sat.	22.785	1.3852	39.356	1.73814	46.609	1.84934	15.67	1.35576	34.172	2.09059
95% Sat.	21.06	1.27082	38.318	1.83668	45.046	1.96297	14.501	1.29256	33.162	2.11146
96% Sat.	20.629	1.32216	37.856	1.81179	44.23	1.98245	14.359	1.27988	32.857	2.13655
97% Sat.	19.926	1.68196	37.42	1.87572	43.571	2.1874	14.134	1.1984	32.688	2.12104
98% Sat.	18.979	1.815	36.852	1.99201	42.254	2.19743	14.076	1.17353	32.105	2.12854
99% Sat.	18.589	1.7212	36.371	1.68855	41.25	1.91574	14.042	1.16719	31.751	2.09949
100% Sat.	18.515	1.71719	36.347	1.68724	41.216	1.91695	14.03	1.16958	31.718	2.09504

Table H.38 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	16.036	2.23836	19.404	2.36397	22.083	3.09223	16.128	2.3246
92% Sat.	15.391	2.12326	18.813	2.28649	21.204	3.13662	15.393	2.10318
93% Sat.	14.96	2.02717	18.065	2.23995	20.353	2.90876	14.923	1.92339
94% Sat.	14.402	2.08498	17.347	2.11427	19.865	2.8701	14.135	1.74622
95% Sat.	13.508	1.61138	15.481	1.421	17.559	2.7004	12.884	1.39612
96% Sat.	12.876	1.32441	14.555	1.54299	16.403	2.50203	12.172	1.39306
97% Sat.	11.963	1.13612	13.564	1.23169	14.855	2.10784	11.299	1.46357
98% Sat.	11.848	1.03137	12.943	1.29719	13.995	1.90032	10.801	1.30615
99% Sat.	11.705	1.01444	12.632	0.97243	13.305	1.51543	10.728	1.17276
100% Sat.	11.704	1.01443	12.629	0.97245	13.284	1.51217	10.725	1.17292

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	19.15	2.59946	19.603	3.69448	21.202	2.63268	17.588	2.50431	19.305	2.45839
92% Sat.	18.319	2.49451	18.841	3.69622	20.172	2.66668	16.81	2.53163	18.608	2.44068
93% Sat.	17.519	2.06668	18.309	3.44221	19.385	2.42606	16.102	2.27634	17.852	2.24357
94% Sat.	16.899	2.10263	17.962	3.45608	18.4	2.47398	15.231	2.13439	17.154	2.1794
95% Sat.	14.923	1.86719	16.246	2.83296	16.257	1.91767	13.666	1.61132	15.283	1.3858
96% Sat.	13.988	1.79363	15.387	2.49574	15.253	1.90443	12.722	1.43136	14.329	1.42644
97% Sat.	12.76	1.41185	14.279	1.97421	14.07	1.38805	11.948	1.28849	13.384	1.15353
98% Sat.	12.605	1.29403	13.772	2.11713	13.405	1.4758	11.591	1.23711	12.666	1.27476
99% Sat.	12.054	1.12336	13.145	1.7705	12.856	0.97061	11.405	1.02993	12.42	1.03886
100% Sat.	12.037	1.10031	13.045	1.61381	12.845	0.96565	11.402	1.02938	12.416	1.0387

Table H.39 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.154	0.07774	17.252	1.64971	7.571	1.24968	1.122	0.36776
92% Sat.	0.13	0.08393	16.679	1.5836	6.628	1.15288	0.932	0.28801
93% Sat.	0.13	0.08393	16.329	1.5603	6.602	1.15712	0.928	0.2838
94% Sat.	0.13	0.08393	15.666	1.47748	6.449	1.10949	0.927	0.284
95% Sat.	0.036	0.02764	12.442	1.107	2.893	0.67103	0.309	0.14221
96% Sat.	0.025	0.02264	11.217	1.21815	2.223	0.46812	0.245	0.10916
97% Sat.	0.021	0.02336	10.255	1.03378	1.872	0.43119	0.205	0.09772
98% Sat.	0.02	0.02309	9.478	1.13492	1.863	0.42797	0.204	0.09762
99% Sat.	0.02	0.02278	9.304	0.9541	1.862	0.42826	0.204	0.09727
100% Sat.	0.019	0.02291	9.303	0.95392	1.861	0.42814	0.203	0.09726

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	1.661	0.37471	16.751	2.1051	18.619	1.52124	11.327	1.16751	17.467	1.66928
92% Sat.	1.381	0.3436	16.189	2.01831	17.733	1.61761	11.16	1.05873	16.805	1.7002
93% Sat.	1.373	0.33966	16.108	2.02318	17.396	1.58032	11.153	1.05818	16.57	1.67014
94% Sat.	1.373	0.33966	15.774	2.00352	16.502	1.55429	11.148	1.05474	15.949	1.67037
95% Sat.	0.527	0.16637	13.939	1.52775	14.336	1.42096	10.608	0.99686	14.129	1.27039
96% Sat.	0.391	0.13352	13.103	1.2997	13.318	1.44452	10.534	1.01249	13.259	1.27722
97% Sat.	0.326	0.11661	12.472	1.09291	12.726	1.23948	10.451	0.99971	12.468	1.16299
98% Sat.	0.325	0.11591	11.913	1.26796	12.025	1.31567	10.41	0.97802	11.718	1.20065
99% Sat.	0.324	0.11542	11.772	1.10171	11.701	0.98592	10.395	0.9645	11.557	1.0081
100% Sat.	0.323	0.11564	11.771	1.10122	11.7	0.98583	10.394	0.96445	11.556	1.00777

Table H.40 Percentages of TTIs (Mean and STD) in the exploitation stage for the HoL Delay satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	1.042	0.43434	47.02	1.37097	16.533	1.62311	3.628	0.90691
92% Sat.	0.934	0.34762	46.701	1.3051	14.733	1.63723	3.336	0.81088
93% Sat.	0.89	0.35459	46.25	1.33808	14.216	1.5654	3.251	0.80549
94% Sat.	0.848	0.36001	45.725	1.28818	13.721	1.50182	3.202	0.80821
95% Sat.	0.416	0.19554	41.449	1.53048	7.611	1.07429	1.276	0.55844
96% Sat.	0.372	0.19416	39.421	1.35137	6.432	0.94917	1.094	0.50153
97% Sat.	0.298	0.1336	37.343	1.84454	5.195	0.88671	0.935	0.37508
98% Sat.	0.263	0.13714	34.879	2.26694	4.866	0.88952	0.894	0.38133
99% Sat.	0.209	0.12264	33.086	2.20233	4.828	0.89337	0.842	0.37842
100% Sat.	0.173	0.12284	33.034	2.19889	4.774	0.89367	0.794	0.37746

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	19.815	1.61896	4.199	1.01997	4.102	0.96705	18.512	1.21884	47.075	1.37161
92% Sat.	18.217	1.46017	3.865	0.93291	3.752	0.83531	16.705	1.22804	46.768	1.28448
93% Sat.	17.727	1.42456	3.774	0.91261	3.674	0.83106	16.23	1.15986	46.307	1.33225
94% Sat.	17.377	1.33806	3.714	0.91186	3.617	0.83385	15.723	1.05633	45.793	1.30751
95% Sat.	13.716	1.20232	1.543	0.59807	1.476	0.62081	10.918	0.76908	41.548	1.58439
96% Sat.	12.914	1.1072	1.325	0.54507	1.263	0.56945	9.87	0.75884	39.562	1.42638
97% Sat.	11.799	1.23664	1.153	0.42276	1.071	0.42568	8.648	0.98157	37.468	1.904
98% Sat.	11.581	1.17387	1.108	0.43062	1.033	0.42946	8.35	0.99185	35.018	2.32298
99% Sat.	11.553	1.16991	1.063	0.4211	0.984	0.41901	8.325	0.98983	33.212	2.25306
100% Sat.	11.526	1.16963	1.002	0.43087	0.928	0.42542	8.288	0.99237	33.157	2.24618

Table H.41 Percentages of TTIs (Mean and STD) in the exploitation stage for the PDR satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	9.693	2.42458	12.896	2.29236	14.882	3.48348	9.59	1.37845
92% Sat.	9.496	2.25725	12.347	2.0245	14.105	3.16498	9.118	1.00045
93% Sat.	9.204	2.07115	11.956	1.9276	13.569	2.77001	8.902	0.90811
94% Sat.	8.725	2.26286	11.483	1.99357	13.321	2.93068	8.45	1.03748
95% Sat.	7.709	1.85689	9.523	1.84222	11.343	2.92785	6.734	0.98168
96% Sat.	6.993	1.61142	8.738	1.53249	10.398	2.82412	6.193	1.10884
97% Sat.	6.588	1.27749	8.454	1.25675	9.462	2.20525	6.071	1.0928
98% Sat.	6.29	0.92659	7.895	1.14225	8.687	2.06119	5.482	0.77277
99% Sat.	6.288	0.92459	7.852	1.13954	8.392	1.73217	5.479	0.77268
100% Sat.	6.288	0.92467	7.848	1.1397	8.39	1.73183	5.477	0.77269

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	14.369	3.19484	10.451	3.3741	10.866	2.22099	14.409	3.40381	13.001	2.52031
92% Sat.	13.735	2.86747	9.992	3.05942	10.22	1.95713	13.811	3.08276	12.546	2.24648
93% Sat.	13.49	2.86832	9.827	2.69615	9.864	1.86498	13.481	2.87767	12.173	2.14193
94% Sat.	13.094	2.99957	9.492	2.77598	9.402	1.98155	13.266	3.05059	11.592	2.2128
95% Sat.	10.291	2.96669	7.746	2.3103	7.42	1.74645	11.113	3.02464	9.408	2.14463
96% Sat.	9.303	2.87683	6.803	1.80618	6.767	1.50005	10.341	2.89759	8.824	2.11672
97% Sat.	8.443	2.24196	6.285	1.62324	6.574	1.37066	9.534	2.4256	8.325	1.71963
98% Sat.	7.7	2.03034	5.95	1.17933	6.015	1.05556	8.826	2.16799	7.783	1.63595
99% Sat.	7.366	1.67377	5.944	1.16965	6.01	1.05269	8.453	1.71504	7.625	1.44736
100% Sat.	7.364	1.67406	5.941	1.16949	6.008	1.05318	8.45	1.71447	7.621	1.44679

Table H.42 Percentages of TTIs (Mean and STD) in the exploitation stage for the DP Multi-Objective satisfaction levels based on PDR-MMF with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: VBR.

Policies / DP Satisfaction	GPF-EDF [Mean]	GPF-EDF [STD]	GPF-LOG [Mean]	GPF-LOG [STD]	GPF-EXP1 [Mean]	GPF-EXP1 [STD]	GPF-EXP2 [Mean]	GPF-EXP2 [STD]
91% Sat.	0.192	0.16747	11.756	1.92832	5.662	1.06675	0.898	0.30024
92% Sat.	0.189	0.16813	11.202	1.66801	5.198	0.91406	0.891	0.30382
93% Sat.	0.189	0.16813	11.108	1.6362	5.111	0.78602	0.89	0.30397
94% Sat.	0.189	0.16813	10.587	1.63629	4.967	0.77245	0.875	0.3087
95% Sat.	0.042	0.05255	7.512	1.48182	1.909	0.41974	0.167	0.11346
96% Sat.	0.034	0.05016	6.726	1.15091	1.431	0.27706	0.142	0.11222
97% Sat.	0.033	0.05026	6.524	1.03565	1.36	0.30471	0.142	0.11207
98% Sat.	0.032	0.05043	6.007	0.94452	1.273	0.23975	0.139	0.11328
99% Sat.	0.032	0.05014	5.969	0.92415	1.272	0.23966	0.139	0.11311
100% Sat.	0.032	0.05014	5.968	0.92385	1.271	0.23975	0.139	0.1128

Policies / DP Satisfaction	**QV [Mean]**	**QV [STD]**	**QV2 [Mean]**	**QV2 [STD]**	**QVMAX [Mean]**	**QVMAX [STD]**	**QVMAX2 [Mean]**	**QVMAX2 [STD]**	**ACLA- [Mean]**	**ACLA [STD]**
91% Sat.	8.901	1.69595	1.192	0.37974	1.136	0.36041	8.199	1.20712	11.824	2.11682
92% Sat.	8.613	1.65788	1.183	0.38535	1.128	0.36226	7.768	1.02641	11.366	1.84859
93% Sat.	8.612	1.65669	1.183	0.38508	1.127	0.36235	7.744	1.00669	11.278	1.8016
94% Sat.	8.416	1.78459	1.16	0.34747	1.114	0.35964	7.594	0.97608	10.68	1.77492
95% Sat.	7.525	1.60811	0.286	0.13684	0.21	0.12695	5.693	0.92398	7.518	1.75378
96% Sat.	7.155	1.67752	0.253	0.12586	0.174	0.11941	5.35	0.88019	6.931	1.68181
97% Sat.	6.939	1.65216	0.252	0.12528	0.173	0.11971	5.282	0.85723	6.538	1.42903
98% Sat.	6.286	1.43617	0.246	0.12288	0.172	0.12021	5.183	0.80038	6.024	1.39429
99% Sat.	6.283	1.42983	0.246	0.12292	0.172	0.12016	5.183	0.80029	5.888	1.19169
100% Sat.	6.282	1.42929	0.245	0.12266	0.171	0.12004	5.182	0.80004	5.887	1.1912

H.3 Percentages of TTIs for the DP Testing Rewards Based on the CBR and VBR Traffic Types

In LTE scheduling, the data packets that have the HoL delays greater than the HoL delay requirements are automatically dropped and declared lost. In this sense, the reward function for the delay parameter should use a fraction from the 3GPP delay requirement as a new delay target as shown in Equation 7.7 from Chapter 7. Other major problem refers to the very high traffic load scheduling under very restrictive HoL delay requirements. In this case, the end of the episode when the HoL delay objective is considered may not be reached. Then, the reward function must use the relaxation parameter $\kappa_D \in \mathbb{R}_{[0,1]}$ as indicated in Equation 7.15. The number of DP feasible states is much lower than the number of episodes due to this parameter. This is beneficial since the closer these values are, the lower the mean HoL delay is for the obtained policies. When the PDR performance is considered, an important role is determined by the windowing factor which is used in the drop rate computations. When the windowing factor is very large, then the PDR objective is degraded and when the windowing factor is very restrictive, the PDR approaches to zero since the dropped packets are not detected during few TTIs. For these simulations, the following windowing factors are used for the DP and PDR reward computations: $\rho = \{5.5, 50, 100, 200, 300, 400, 500\}$. The same policies from Section H.2 are evaluated in terms of the mean percentage of TTIs when the rewards are moderate, punishment and maximized for the HoL delay, PDR and DP objectives. The rest of this section is organized as follows: Tables H.43 to H. 63 present the simulation results of the reward types when the CBR traffic is used and Tables H.64 to H.84 indicate the amount of different types of rewards for the VBR traffic type. The obtained scheduling policies are declared sustainable from the viewpoint of the DP multi-objective evaluation if: the mean percentage of DP feasible TTIs is maximized, the STD values and the amount of punishment rewards are minimized, and finally, the discrepancy between the number of DP feasible states and the number of episodes $\left(\mathcal{RW}_t^{DP} = 1\right)$ is minimized.

Table H.43 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^D = 1$, $\mathcal{RW}_t^D > 0$ and $\mathcal{RW}_t^D < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^D < 0$	Punish Reward STD[%] $\mathcal{RW}_t^D < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^D > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^D > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^D = 1$	Max. Reward STD[%] $\mathcal{RW}_t^D = 1$
QV	0.141	0	27.662	0.9741	72.196	0.9741
QV2	0.141	0	27.319	0.88969	72.538	0.88969
QVMAX	0.141	0	26.683	0.9922	73.175	0.9922
QVMAX2	0.141	0	27.389	1.01546	72.468	1.01546
ACLA	0.141	0	25.933	1.26389	73.925	1.26389
Best			25.933	**ACLA**	73.925	**ACLA**
Worst			27.662	**QV**	72.196	**QV**

Table H.44 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme: $(Top3, N_{CT} = 64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	9.72	0.39359	3.402	0.45315	86.876	0.54529
QV2	9.515	0.35706	4.276	0.45024	86.208	0.55343
QVMAX	9.096	0.14947	6.387	0.68306	84.516	0.74602
QVMAX2	9.09	0.14558	4.822	0.81461	86.088	0.89713
ACLA	9.402	0.16514	6.945	0.71116	83.651	0.79877
Best	9.09	**QVMAX2**	3.402	**QV**	86.876	**QV**
Worst	9.72	**QV**	6.945	**ACLA**	83.651	**ACLA**

Table H.45 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP}=1$, $\mathcal{RW}_t^{DP}>0$ and $\mathcal{RW}_t^{DP}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP}=1$
QV	20.026	0.90527	9.179	0.46673	70.794	1.18306
QV2	20.178	0.88353	8.662	0.4292	71.158	1.06908
QVMAX	19.574	0.94429	8.645	0.4205	71.78	1.17648
QVMAX2	19.766	0.92352	9.177	0.52008	71.056	1.20588
ACLA	18.499	1.00984	9.059	0.50991	72.441	1.41791
Best	18.499	**ACLA**	8.645	**QVMAX**	72.441	**ACLA**
Worst	20.178	**QV2**	9.179	**QV**	70.794	**QV**

Table H.46 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{D}=1$, $\mathcal{RW}_t^{D}>0$ and $\mathcal{RW}_t^{D}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=50)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{D}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{D}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{D}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{D}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{D}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{D}=1$
QV	0.141	0	27.246	0.63238	72.612	0.63238
QV2	0.141	0	26.797	1.15891	73.06	1.15891
QVMAX	0.141	0	28.698	0.45783	71.16	0.45783
QVMAX2	0.141	0	32.581	1.31157	67.276	1.31157
ACLA	0.141	0	29.132	0.33293	70.725	0.33293
Best			26.797	**QV2**	73.06	**QV2**
Worst			32.581	**QVMAX2**	67.276	**QVMAX2**

Table H.47 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	12.93	0.21145	15.243	0.81614	71.826	0.96166
QV2	14.733	0.68309	8.076	0.72915	77.19	0.98208
QVMAX	12.903	0.26403	14.715	0.51121	72.381	0.72993
QVMAX2	12.905	0.24061	15.841	0.69859	71.253	0.834
ACLA	14.742	0.57929	12.744	0.67661	72.513	0.7793
Best	12.903	**QVMAX**	8.076	**QV2**	77.19	**QV2**
Worst	14.742	**ACLA**	15.841	**QVMAX2**	71.253	**QVMAX2**

Table H.48 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP} = 1$, $\mathcal{RW}_t^{DP} > 0$ and $\mathcal{RW}_t^{DP} < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP} < 0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP} < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP} > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP} > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP} = 1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP} = 1$
QV	19.158	0.69046	17.14	0.52429	63.701	0.56349
QV2	19.495	0.93322	15.388	0.48135	65.116	0.93038
QVMAX	21.044	0.47364	15.606	0.50691	63.349	0.42861
QVMAX2	23.163	0.74866	17.059	0.55566	59.777	0.99772
ACLA	21.218	0.41794	15.409	0.60931	63.372	0.39736
Best	19.158	**QV**	15.388	**QV2**	65.116	**QV2**
Worst	23.163	**QVMAX2**	17.14	**QV**	59.777	**QVMAX2**

Table H.49 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^D = 1$, $\mathcal{RW}_t^D > 0$ and $\mathcal{RW}_t^D < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^D < 0$	Punish Reward STD[%] $\mathcal{RW}_t^D < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^D > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^D > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^D = 1$	Max. Reward STD[%] $\mathcal{RW}_t^D = 1$
QV	0.141	0	29.552	0.6185	70.306	0.6185
QV2	0.141	0	27.497	0.65221	72.361	0.65221
QVMAX	0.141	0	29.911	0.86274	69.946	0.86274
QVMAX2	0.141	0	28.218	0.8187	71.64	0.8187
ACLA	0.141	0	27.008	0.67384	72.849	0.67384
Best			27.008	**ACLA**	72.849	**ACLA**
Worst			29.911	**QVMAX**	69.946	**QVMAX**

Table H.50 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^D > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^D > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^G = 1$	Max. Reward STD[%] $\mathcal{RW}_t^G = 1$
QV	15.986	0.6012	23.43	1.34103	60.583	1.59599
QV2	16.254	0.49362	21.21	1.17375	62.535	1.27157
QVMAX	18.247	0.97604	15.892	1.43222	65.86	1.3764
QVMAX2	17.359	0.74724	15.377	1.40295	67.262	1.65412
ACLA	17.225	0.78622	16.444	1.50897	66.33	1.73523
Best	15.986	**QV**	15.377	**QVMAX2**	67.262	**QVMAX2**
Worst	18.247	**QVMAX**	23.43	**QV**	60.583	**QV**

Table H.51 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP}=1$, $\mathcal{RW}_t^{DP}>0$ and $\mathcal{RW}_t^{DP}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=100)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP}=1$
QV	20.659	0.64791	24.284	0.65571	55.056	0.96928
QV2	20.023	0.62563	23.172	0.86308	56.803	1.216
QVMAX	21.468	0.71567	21.428	1.13588	57.103	1.55039
QVMAX2	20.264	0.79312	21.542	0.88871	58.193	1.39006
ACLA	19.747	0.68175	21.915	1.18863	58.336	1.52013
Best	19.747	**ACLA**	21.428	**QVMAX**	58.336	**ACLA**
Worst	21.468	**QVMAX**	24.284	**QV**	55.056	**QV**

Table H.52 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{D}=1$, $\mathcal{RW}_t^{D}>0$ and $\mathcal{RW}_t^{D}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=200)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{D}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{D}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{D}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{D}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{D}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{D}=1$
QV	0.141	0	29.296	0.79846	70.562	0.79846
QV2	0.141	0	32.808	0.91255	67.049	0.91255
QVMAX	0.141	0	28.576	0.26703	71.282	0.26703
QVMAX2	0.141	0	26.939	0.68301	72.919	0.68301
ACLA	0.141	0	30.745	0.91456	69.113	0.91456
Best			26.939	**QVMAX2**	72.919	**QVMAX2**
Worst			32.808	**QV2**	67.049	**QV2**

Table H.53 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1, \mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$; Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	24.076	0.68158	29.256	1.56494	46.667	1.50838
QV2	22.614	0.67134	33.503	1.31706	43.882	1.09978
QVMAX	22.381	0.58869	32.274	1.118	45.345	1.45115
QVMAX2	24.254	0.64315	26.793	1.00861	48.952	0.68799
ACLA	23.025	0.57059	32.833	1.09264	44.141	1.07391
Best	22.381	**QVMAX**	26.793	**QVMAX2**	48.952	**QVMAX2**
Worst	24.254	**QVMAX2**	33.503	**QV2**	43.882	**QV2**

Table H.54 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP} = 1, \mathcal{RW}_t^{DP} > 0$ and $\mathcal{RW}_t^{DP} < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 200)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP} < 0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP} < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP} > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP} > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP} = 1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP} = 1$
QV	20.119	0.54922	35.792	1.62493	44.088	1.4703
QV2	23.33	0.48665	34.634	1.21087	42.034	1.16248
QVMAX	20.883	0.46611	36.148	1.41403	42.968	1.28536
QVMAX2	19.621	0.65566	34.78	1.08647	45.599	0.7785
ACLA	21.945	0.54968	35.459	1.14763	42.595	1.08901
Best	19.621	**QVMAX2**	34.634	**QV2**	45.599	**QVMAX2**
Worst	23.33	**QV2**	36.148	**QVMAX**	42.034	**QV2**

Table H.55 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^D = 1$, $\mathcal{RW}_t^D > 0$ and $\mathcal{RW}_t^D < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^D < 0$	Punish Reward STD[%] $\mathcal{RW}_t^D < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^D > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^D > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^D = 1$	Max. Reward STD[%] $\mathcal{RW}_t^D = 1$
QV	0.141	0	29.123	0.98589	70.735	0.98589
QV2	0.141	0	34.056	1.49975	65.802	1.49975
QVMAX	0.141	0	27.253	0.82497	72.605	0.82497
QVMAX2	0.141	0	34.385	1.7497	65.472	1.7497
ACLA	0.141	0	34.359	1.47005	65.499	1.47005
Best			27.253	**QVMAX**	72.605	**QVMAX**
Worst			34.385	**QVMAX2**	65.472	**QVMAX2**

Table H.56 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	30.354	1.33468	37.139	1.791	32.506	2.05852
QV2	30.324	1.3749	37.388	1.77868	32.286	1.86872
QVMAX	31.912	1.54058	29.728	2.58009	38.359	3.14684
QVMAX2	33.544	1.89486	33.033	2.45997	33.422	3.35869
ACLA	30.957	1.18679	35.851	1.87318	33.191	2.3415
Best	30.324	**QV2**	29.728	**QVMAX**	38.359	**QVMAX**
Worst	33.544	**QVMAX2**	37.388	**QV2**	32.286	**QV2**

Table H.57 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP}=1$, $\mathcal{RW}_t^{DP}>0$ and $\mathcal{RW}_t^{DP}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=300)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP}=1$
QV	20.791	0.72586	48.187	1.92183	31.021	1.78086
QV2	23.921	0.87736	45.522	2.01889	30.556	1.70585
QVMAX	19.83	0.7155	43.336	2.65771	36.833	2.88278
QVMAX2	24.37	1.0743	44.719	3.02081	30.909	3.04784
ACLA	24.184	0.90942	44.254	2.64062	31.561	2.28296
Best	19.83	**QVMAX**	43.336	**QVMAX**	36.833	**QVMAX**
Worst	24.37	**QVMAX2**	48.187	**QV**	30.556	**QV2**

Table H.58 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{D}=1$, $\mathcal{RW}_t^{D}>0$ and $\mathcal{RW}_t^{D}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=400)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{D}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{D}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{D}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{D}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{D}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{D}=1$
QV	0.141	0	28.009	0.66226	71.848	0.66226
QV2	0.141	0	30.06	0.72744	69.798	0.72744
QVMAX	0.141	0	26.08	0.81598	73.777	0.81598
QVMAX2	0.141	0	26.856	0.94288	73.002	0.94288
ACLA	0.141	0	27.055	0.8226	72.803	0.8226
Best			26.08	**QVMAX**	73.777	**QVMAX**
Worst			30.06	**QV2**	69.798	**QV2**

Table H.59 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	38.005	1.29212	31.585	4.21127	30.408	3.96569
QV2	38.788	1.47889	32.014	4.16026	29.197	4.19584
QVMAX	36.887	1.38853	34.062	3.6375	29.05	3.43734
QVMAX2	38.976	1.87033	30.411	4.49422	30.612	4.07156
ACLA	35.211	1.28885	38.197	2.39554	26.591	2.83261
Best	35.211	**ACLA**	30.411	**QVMAX2**	30.612	**QVMAX2**
Worst	38.976	**QVMAX2**	38.197	**ACLA**	26.591	**ACLA**

Table H.60 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP} = 1$, $\mathcal{RW}_t^{DP} > 0$ and $\mathcal{RW}_t^{DP} < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP} < 0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP} < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP} > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP} > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP} = 1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP} = 1$
QV	20.113	0.68568	50.509	3.85596	29.377	3.91228
QV2	21.25	0.6393	50.513	3.82694	28.236	4.06305
QVMAX	19.137	0.70542	52.519	3.30009	28.343	3.4682
QVMAX2	19.554	0.75401	50.895	4.05076	29.549	4.00352
ACLA	18.974	0.746	55.145	2.55506	25.88	2.74268
Best	18.974	**ACLA**	50.509	**QV**	29.549	**QVMAX2**
Worst	21.25	**QV2**	55.145	**ACLA**	25.88	**ACLA**

Table H.61 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^D = 1$, $\mathcal{RW}_t^D > 0$ and $\mathcal{RW}_t^D < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^D < 0$	Punish Reward STD[%] $\mathcal{RW}_t^D < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^D > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^D > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^D = 1$	Max. Reward STD[%] $\mathcal{RW}_t^D = 1$
QV	0.141	0	29.624	1.0143	70.233	1.0143
QV2	0.141	0	29.211	0.72226	70.646	0.72226
QVMAX	0.141	0	25.999	0.73334	73.859	0.73334
QVMAX2	0.141	0	33.777	0.92171	66.081	0.92171
ACLA	0.141	0	28.293	0.66365	71.564	0.66365
Best			25.999	**QVMAX**	73.859	**QVMAX**
Worst			33.777	**QVMAX2**	66.081	**QVMAX2**

Table H.62 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	45.221	2.6032	32.074	3.40722	22.704	2.68689
QV2	44.076	2.97269	29.72	3.14087	26.203	0.87661
QVMAX	41.805	1.57511	34.434	3.54136	23.76	3.03763
QVMAX2	44.412	1.6836	36.667	2.27675	18.92	2.22902
ACLA	41.858	1.80856	37.133	2.29799	21.008	2.78537
Best	41.805	**QVMAX**	29.72	**QV2**	26.203	**QV2**
Worst	45.221	**QV**	37.133	**ACLA**	18.92	**QVMAX2**

Table H.63 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP}=1$, $\mathcal{RW}_t^{DP}>0$ and $\mathcal{RW}_t^{DP}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=500)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: CBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP}=1$
QV	21.457	0.71744	57.144	2.32082	21.398	2.63943
QV2	21.026	0.5201	53.785	0.94627	25.187	1.0029
QVMAX	18.656	0.59309	58.081	2.70816	23.261	3.05788
QVMAX2	23.945	0.50798	57.636	2.01913	18.418	2.15596
ACLA	20.144	0.4989	59.395	2.27274	20.459	2.62503
Best	18.656	**QVMAX**	53.785	**QV2**	25.187	**QV2**
Worst	23.945	**QVMAX2**	59.395	**ACLA**	18.418	**QVMAX2**

Table H.64 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{D}=1$, $\mathcal{RW}_t^{D}>0$ and $\mathcal{RW}_t^{D}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{D}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{D}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{D}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{D}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{D}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{D}=1$
QV	0.142	0.00047	52.742	2.3474	47.114	2.34744
QV2	0.142	0	67.427	1.80881	32.429	1.80881
QVMAX	0.142	0.00059	67.447	2.15412	32.41	2.15401
QVMAX2	0.142	0.00054	52.136	1.64714	47.721	1.64704
ACLA	0.142	0.00054	47.338	1.64857	52.52	1.64832
Best			47.338	**ACLA**	52.52	**ACLA**
Worst			67.447	**QVMAX**	32.41	**QVMAX**

Table H.65 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	13.581	1.13772	9.526	0.55142	76.891	1.25799
QV2	19.042	1.47499	11.707	1.12983	69.249	1.72178
QVMAX	11.411	1.04415	21.256	1.34702	67.332	1.59153
QVMAX2	11.467	1.19898	17.241	1.09153	71.29	1.44831
ACLA	10.208	0.85019	22.358	1.47834	67.433	1.7757
Best	10.208	**ACLA**	9.526	**QV**	76.891	**QV**
Worst	19.042	**QV2**	22.358	**ACLA**	67.332	**QVMAX**

Table H.66 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP} = 1$, $\mathcal{RW}_t^{DP} > 0$ and $\mathcal{RW}_t^{DP} < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 5.5)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP} < 0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP} < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP} > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP} > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP} = 1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP} = 1$
QV	35.895	1.31927	17.945	1.05672	46.159	2.28565
QV2	43.386	0.85089	24.959	1.05681	31.654	1.80351
QVMAX	43.881	1.10007	24.459	1.15276	31.659	2.16578
QVMAX2	35.552	0.96097	17.677	0.8544	46.77	1.52991
ACLA	32.469	1.15057	16.476	0.56868	51.055	1.45753
Best	32.469	**ACLA**	16.476	**ACLA**	51.055	**ACLA**
Worst	43.881	**QVMAX**	24.959	**QV2**	31.654	**QV2**

Table H.67 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^D = 1$, $\mathcal{RW}_t^D > 0$ and $\mathcal{RW}_t^D < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^D < 0$	Punish Reward STD[%] $\mathcal{RW}_t^D < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^D > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^D > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^D = 1$	Max. Reward STD[%] $\mathcal{RW}_t^D = 1$
QV	0.142	0.00059	50.455	1.66677	49.402	1.6666
QV2	0.142	0	52.538	1.96371	47.319	1.96371
QVMAX	0.142	0.00054	52.056	2.15658	47.801	2.15645
QVMAX2	0.142	0	51.174	2.31165	48.682	2.31165
ACLA	0.141	0	51.894	2.10713	47.964	2.10713
Best			50.455	**QV**	49.402	**QV**
Worst			52.538	**QV2**	47.319	**QV2**

Table H.68 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 50)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	18.206	1.85887	24.908	1.74724	56.884	2.14918
QV2	27.065	2.67029	10.899	0.95434	62.035	3.16735
QVMAX	21.646	1.68401	13.671	1.73516	64.682	2.64413
QVMAX2	26.679	2.4407	8.874	0.87665	64.446	2.95659
ACLA	16.714	1.88031	28.953	1.98473	54.332	1.94992
Best	16.714	**ACLA**	8.874	**QVMAX2**	64.682	**QVMAX**
Worst	27.065	**QV2**	28.953	**ACLA**	54.332	**ACLA**

Table H.69 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP}=1$, $\mathcal{RW}_t^{DP}>0$ and $\mathcal{RW}_t^{DP}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=50)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP}=1$
QV	34.33	0.86487	21.51	0.88025	44.159	1.6119
QV2	35.745	1.15349	22.396	0.86373	41.858	1.91196
QVMAX	35.114	1.07564	22.004	1.09673	42.881	1.9654
QVMAX2	35.279	1.34629	21.668	0.97085	43.051	2.11515
ACLA	34.98	1.14086	22.442	0.91333	42.578	1.85367
Best	34.33	**QV**	21.51	**QV**	44.159	**QV**
Worst	35.745	**QV2**	22.442	**ACLA**	41.858	**QV2**

Table H.70 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{D}=1$, $\mathcal{RW}_t^{D}>0$ and $\mathcal{RW}_t^{D}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=100)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{D}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{D}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{D}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{D}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{D}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{D}=1$
QV	0.142	0.00058	54.898	1.42146	44.959	1.42142
QV2	0.142	0.00047	53.42	2.03702	46.437	2.03702
QVMAX	0.142	0.00054	49.364	2.31464	50.493	2.31465
QVMAX2	0.142	0	48.7	2.34333	51.157	2.34333
ACLA	0.142	0	49.185	2.34396	50.672	2.34396
Best			48.7	**QVMAX2**	51.157	**QVMAX2**
Worst			54.898	**QV**	44.959	**QV**

Table H.71 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1, \mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	24.068	2.01908	30.837	1.72995	45.094	2.41685
QV2	25.38	1.96287	26.341	2.93787	48.278	3.67953
QVMAX	22.365	1.82168	31.955	2.0988	45.679	2.64714
QVMAX2	29.778	2.08291	17.946	1.4086	52.275	2.53269
ACLA	28.885	2.1442	18.438	1.31237	52.676	2.64178
Best	22.365	**QVMAX**	17.946	**QVMAX2**	52.676	**ACLA**
Worst	29.778	**QVMAX2**	31.955	**QVMAX**	45.094	**QV**

Table H.72 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP} = 1$, $\mathcal{RW}_t^{DP} > 0$ and $\mathcal{RW}_t^{DP} < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 100)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP} < 0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP} < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP} > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP} > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP} = 1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP} = 1$
QV	37.194	0.99335	27.531	0.52279	35.274	1.1239
QV2	35.797	1.5356	25.232	0.80349	38.969	2.09919
QVMAX	33.801	1.75531	26.555	0.45512	39.642	1.85064
QVMAX2	33.322	1.59635	26.133	0.85938	40.543	1.97066
ACLA	33.187	1.51635	26.551	1.04185	40.261	2.12226
Best	33.187	**ACLA**	25.232	**QV2**	40.543	**QVMAX2**
Worst	37.194	**QV**	27.531	**QV**	35.274	**QV**

Table H.73 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{\mathcal{D}}=1$, $\mathcal{RW}_t^{\mathcal{D}}>0$ and $\mathcal{RW}_t^{\mathcal{D}}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=200)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{\mathcal{D}}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{\mathcal{D}}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{\mathcal{D}}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{\mathcal{D}}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{\mathcal{D}}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{\mathcal{D}}=1$
QV	0.142	0.00059	47.25	1.7533	52.607	1.75325
QV2	0.142	0.00035	50.15	1.99475	49.706	1.99464
QVMAX	0.142	0.00059	48.664	1.92533	51.193	1.92531
QVMAX2	0.142	0	50.155	1.99585	49.702	1.99585
ACLA	0.142	0.00058	49.583	1.9803	50.274	1.9803
Best			47.25	**QV**	52.607	**QV**
Worst			50.155	**QVMAX2**	49.702	**QVMAX2**

Table H.74 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{\mathcal{P}}=1$, $\mathcal{RW}_t^{\mathcal{P}}>0$ and $\mathcal{RW}_t^{\mathcal{P}}<0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=200)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{\mathcal{P}}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{\mathcal{P}}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{\mathcal{P}}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{\mathcal{P}}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{\mathcal{P}}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{\mathcal{P}}=1$
QV	30.485	3.37669	39.764	3.12921	29.75	0.98396
QV2	39.706	3.40543	26.642	2.26011	33.651	2.18168
QVMAX	31.82	3.01097	37.03	2.86717	31.149	1.53512
QVMAX2	41.089	3.41351	25.28	2.1134	33.63	2.91093
ACLA	30.843	3.52011	39.865	3.40437	29.291	0.79845
Best	30.485	**QV**	25.28	**QVMAX2**	33.651	**QV2**
Worst	41.089	**QVMAX2**	39.865	**ACLA**	29.291	**ACLA**

Table H.75 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP}=1$, $\mathcal{RW}_t^{DP}>0$ and $\mathcal{RW}_t^{DP}<0$. DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=200)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP}=1$
QV	32.194	1.21978	38.545	1.10094	29.26	0.91248
QV2	34.047	1.2525	35.594	1.23817	30.357	1.5529
QVMAX	33.545	1.29079	36.504	1.05289	29.95	1.03631
QVMAX2	34.094	1.28783	35.598	1.64145	30.307	2.06153
ACLA	33.646	1.27245	38.514	0.89162	27.839	0.98382
Best	32.194	**QV**	35.594	**QV2**	30.357	**QV2**
Worst	34.094	**QVMAX2**	38.545	**QV**	27.839	**ACLA**

Table H.76 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{D}=1$, $\mathcal{RW}_t^{D}>0$ and $\mathcal{RW}_t^{D}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=300)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{D}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{D}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{D}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{D}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{D}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{D}=1$
QV	0.142	0.00058	47.254	1.75622	52.603	1.75648
QV2	0.142	0.00054	47.726	1.95766	52.132	1.95765
QVMAX	0.142	0.00047	54.534	2.24527	45.322	2.24562
QVMAX2	0.142	0.00059	65.676	2.30745	34.181	2.30737
ACLA	0.141	0.00047	46.756	1.66116	53.101	1.66136
Best			46.756	**ACLA**	53.101	**ACLA**
Worst			65.676	**QVMAX2**	34.181	**QVMAX2**

Table H.77 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	43.403	4.03704	36.63	3.81746	19.965	1.54405
QV2	41.361	4.32253	37.687	4.17187	20.951	1.67752
QVMAX	54.641	2.86033	24.676	2.43267	20.681	2.21422
QVMAX2	45.959	4.55124	35.606	4.19335	18.435	1.63342
ACLA	39.394	3.91619	40.501	3.95392	20.104	1.61646
Best	39.394	**ACLA**	24.676	**QVMAX**	20.951	**QV2**
Worst	54.641	**QVMAX**	40.501	**ACLA**	18.435	**QVMAX2**

Table H.78 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP} = 1$, $\mathcal{RW}_t^{DP} > 0$ and $\mathcal{RW}_t^{DP} < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 300)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP} < 0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP} < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP} > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP} > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP} = 1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP} = 1$
QV	31.469	1.25037	49.135	1.54571	19.395	1.25512
QV2	32.93	1.28221	46.741	1.54463	20.328	1.35437
QVMAX	36.82	1.34068	43.394	2.23012	19.785	1.98374
QVMAX2	42.455	1.22559	41.995	0.97472	15.549	1.81227
ACLA	31.821	1.1988	48.508	1.55466	19.67	1.42428
Best	31.469	**QV**	41.995	**QVMAX2**	20.328	**QV2**
Worst	42.455	**QVMAX2**	49.135	**QV**	15.549	**QVMAX2**

Table H.79 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^D = 1$, $\mathcal{RW}_t^D > 0$ and $\mathcal{RW}_t^D < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^D < 0$	Punish Reward STD[%] $\mathcal{RW}_t^D < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^D > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^D > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^D = 1$	Max. Reward STD[%] $\mathcal{RW}_t^D = 1$
QV	0.142	0	55.334	1.01763	44.522	1.01763
QV2	0.142	0.00054	54.265	2.28476	45.593	2.28433
QVMAX	0.142	0.00058	48.212	1.55429	51.644	1.55402
QVMAX2	0.142	0.00058	64.661	2.41649	35.196	2.41622
ACLA	0.142	0.00054	54.423	2.34382	45.434	2.34361
Best			48.212	**QVMAX**	51.644	**QVMAX**
Worst			64.661	**QVMAX2**	35.196	**QVMAX2**

Table H.80 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 400)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	61.622	4.25411	26.278	3.37249	12.099	1.1023
QV2	59.246	4.53421	27.642	3.29533	13.111	1.61453
QVMAX	51.6	3.92593	35.487	3.3243	12.911	0.96696
QVMAX2	52.818	3.73533	35.715	3.25084	11.466	1.03115
ACLA	47.99	3.23729	39.527	3.02873	12.482	1.0405
Best	47.99	**ACLA**	26.278	**QV**	13.111	**QV2**
Worst	61.622	**QV**	39.527	**ACLA**	11.466	**QVMAX2**

Table H.81 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP}=1$, $\mathcal{RW}_t^{DP}>0$ and $\mathcal{RW}_t^{DP}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=400)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP}=1$
QV	37.586	0.55958	51.863	0.97923	10.55	0.89727
QV2	36.824	1.24662	50.551	1.02171	12.624	1.35732
QVMAX	33.16	1.09717	54.04	1.15299	12.798	0.96538
QVMAX2	42.198	1.09854	46.475	0.93064	11.325	1.03015
ACLA	35.786	1.25181	51.841	0.9115	12.372	1.04427
Best	33.16	**QVMAX**	46.475	**QVMAX2**	12.798	**QVMAX**
Worst	42.198	**QVMAX2**	54.04	**QVMAX**	10.55	**QV**

Table H.82 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{D}=1$, $\mathcal{RW}_t^{D}>0$ and $\mathcal{RW}_t^{D}<0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho=500)$ and CQI Aggregation Scheme $(Top3, N_{CT}=64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{D}<0$	Punish Reward STD[%] $\mathcal{RW}_t^{D}<0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{D}>0$	Moderate Reward STD[%] $\mathcal{RW}_t^{D}>0$	Max. Reward Mean[%] $\mathcal{RW}_t^{D}=1$	Max. Reward STD[%] $\mathcal{RW}_t^{D}=1$
QV	0.142	0.00054	52.195	1.81056	47.662	1.81058
QV2	0.142	0	66.174	2.01804	33.683	2.01804
QVMAX	0.142	0.00058	66.53	2.13776	33.327	2.13768
QVMAX2	0.142	0.00054	51.673	1.46451	48.184	1.46454
ACLA	0.141	0.00047	46.692	1.46014	53.166	1.45995
Best			46.692	**ACLA**	53.166	**ACLA**
Worst			66.53	**QVMAX**	33.327	**QVMAX**

Table H.83 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^P = 1$, $\mathcal{RW}_t^P > 0$ and $\mathcal{RW}_t^P < 0$. The PDR Requirement is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^P < 0$	Punish Reward STD[%] $\mathcal{RW}_t^P < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^P > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^P > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^P = 1$	Max. Reward STD[%] $\mathcal{RW}_t^P = 1$
QV	65.582	3.51205	27.001	3.03595	7.416	1.67441
QV2	71.79	2.79522	22.22	2.12133	5.99	1.17051
QVMAX	60.378	4.2874	33.564	4.26225	6.057	1.05381
QVMAX2	61.46	4.36956	30.037	3.96794	8.502	1.71517
ACLA	54.789	5.1868	37.537	4.18381	7.673	1.44742
Best	54.789	**ACLA**	22.22	**QV2**	8.502	**QVMAX2**
Worst	71.79	**QV2**	37.537	**ACLA**	5.99	**QV2**

Table H.84 Percentages of TTIs (Mean and STD) in the exploitation stage when the testing rewards are: $\mathcal{RW}_t^{DP} = 1$, $\mathcal{RW}_t^{DP} > 0$ and $\mathcal{RW}_t^{DP} < 0$. The DP Multi-Objective is evaluated based on NAUT-MMF user rates with static windowing factor $(\rho = 500)$ and CQI Aggregation Scheme $(Top3, N_{CT} = 64)$. Traffic Type: VBR

Reward Type / RL Alg.	Punish Reward Mean[%] $\mathcal{RW}_t^{DP} < 0$	Punish Reward STD[%] $\mathcal{RW}_t^{DP} < 0$	Moderate Reward Mean[%] $\mathcal{RW}_t^{DP} > 0$	Moderate Reward STD[%] $\mathcal{RW}_t^{DP} > 0$	Max. Reward Mean[%] $\mathcal{RW}_t^{DP} = 1$	Max. Reward STD[%] $\mathcal{RW}_t^{DP} = 1$
QV	35.923	1.19153	57.029	1.60344	7.046	1.70311
QV2	43.027	1.18559	51.633	1.24731	5.338	1.03037
QVMAX	43.505	1.26146	51.117	1.45977	5.377	0.76947
QVMAX2	35.908	0.96738	55.755	1.84937	8.336	1.69521
ACLA	32.006	1.18201	32.006	0.85116	7.569	1.44343
Best	32.006	**ACLA**	32.006	**ACLA**	8.336	**QVMAX2**
Worst	43.505	**QVMAX**	57.029	**QV**	5.338	**QV2**

H.4 Summary

The best set of scheduling policies which are able to maximize the mean percentage of DP feasible TTIs for the CBR traffic type is illustrated in Figure 7.8, Chapter 7. Only when the PDR objective is considered for the CBR traffic type, the best mean percentage of feasible TTIs decreases from 86.786% when $\rho = 5.5$, to 23.699% when $\rho = 500$. If the windowing factor is greater than $\rho > 200$ for the CBR traffic type, the STD value for the mean percentage of DP feasible TTIs increases considerably (>3.0). For this reason, the optimum range of windowing factor for the CBR traffic type from the viewpoint of DP feasible TTIs is $\rho \in [5.5, 200]$. When the VBR traffic type is analysed, the best set of policies which are able to maximize the number of DP feasible TTIs is highlighted in Figure 7.16, Chapter 7. The best mean percentage of DP feasible TTIs registers a decrease from 76.793% (when $\rho = 5.5$) to 8.45% (when $\rho = 500$). Moreover, the optimum windowing factor from the viewpoint of $\overline{p}_{TTI}^{DP,x}$, $x = 91\%,..,100\%$ is $\rho \in [5.5, 300]$ since beyond of this interval, the STD values increase considerably for the entire domain of DP evaluation.

The best mean percentage of TTIs with the punishment rewards for the PDR objective increases from 9.09% (when $\rho = 5.5$) to 41.805% (when $\rho = 500$) for the CBR traffic, which means that the optimum windowing factor for the sustainable set of policies is $\rho = 5.5$ (Table H.44). The same optimum value is valid for the VBR traffic type, where the best mean percentage of TTIs for the PDR punishments increases from 10.208% (when $\rho = 5.5$) to 54.789% (when $\rho = 500$). But the windowing factor needs to be determined based on the GBR and NGMN fairness objectives. Therefore, for both traffic types, it is preferable to use a dynamic windowing factor in the interval of $\rho \in [5.5, 200]$ in order to maximize the mean percentage of TTIs when the scheduler is declared feasible from the viewpoint of the set of objectives such as: NGMN fairness requirement, GBR, HoL packet delay and PDR objectives.

As seen in Chapter 6, Appendix F and Appendix G, the windowing factor plays a very important role in the NGMN fairness and GBR objectives when the AUT-MMF observations are computed. As concluded in Section F.4 from Appendix F, the CACLA2 policy is sustainable if the considered windowing factor belongs to $\rho \in [2.5, 4.0]$ for the infinite buffer traffic type. As concluded in Section G.4 from Appendix G, when the same traffic type is simulated for the DSR-SMOO problems focusing on the GBR objective, the optimum windowing factor is $\rho = 5.5$. When the DSR-CMOO problems focusing on both GBR and NGMN fairness are considered, then the optimum range of windowing factor should be reached in order to manage the trade-off between these objectives. If the windowing factor is $\rho > 4.0$, then the NGMN fairness objective is harmed. In this case, the fairness reward starts to fluctuate and the mean percentage of NGMN feasible TTIs starts to decrease. If the windowing factor is $\rho < 5.5$, then the mean percentage of GBR feasible TTIs is deteriorated and the performance of the mean percentage of NGMN feasible TTIs is improved. Therefore, the windowing factor controls the quantity of the mean percentages of feasible TTIs when both GBR and NGMN fairness objectives are considered.

As mentioned earlier, the lower the windowing factor is, the better is the PDR performance. For higher windowing factors, the PDR performance is strongly affected since the number of dropped packets is counted on larger time windows. When the optimum range of the windowing factor $(\rho \in [5.5, 200])$ in the case of the PDR objective is considered, the mean percentage of GBR feasible TTIs can be increased. On the other side, for the NGMN fairness requirement, the feasible area in the CDF domain should be enlarged in order to increase the number of NGMN feasible TTIs when the windowing factor belongs to $\rho \in [5.5, 200]$. To conclude, the windowing factor is one of the most important element when the trade-off between the NGMN fairness, GBR and PDR objectives is analysed. In this sense, CACLA2+ RL approach which is proposed in Chapter 7, aims to adapt the fairness parameters and the windowing factor in an intelligent manner each time when the fairness agent is selected by the QoS agent as shown in the proposed architecture exposed in Chapter 5.

Bibliography

[1] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*, The Prentice Hall Communications Engineering and Emerging Technologies Series, 2010.

[2] Yang, H. 'A Road to Future Broadband Wireless Access: MIMO-OFDM-Based Air Interface,' *IEEE Communications Magazine*, 43 (1) January 2005, pp. 53-60.

[3] R. Trestian, *User-Centric Power-Friendly Quality-based Network Selection Strategy for Heterogeneous Wireless Environments*, Ph.D. Thesis, Dublin City University, 2012.

[4] 3G Americas White Paper, *The Case For Evolved EDGE*, August 2008.

[5] Y. Jiao. 'Understanding EDGE Evolution and its Measurements,' High Frequency Electronics, pp. 44-50, October 2010.

[6] Adachi, F. 'Evolution Towards Broadband Wireless Systems,' *The 5th International Symposium on Wireless Personal Multimedia Communications*, vol. 1, 27-30 Oct. 2002, pp. 19-26.

[7] ITU-R, *Circular letter 5/LCCE/2*, Tech. Rep., March 2008.

[8] A.D. Little, *The Business Benefits of 4G LTE*, 2012.

[9] Swisscom, *Itinérance 4G/LTE pour les clients de Swisscom,* Communiqué de Presse, Bern, 20 June, 2013.

[10] Real Wireless White Paper *LTE and HSPA device availability in UK-relevant frequency bands: Current availability and future evolution,* v. 1.03, May 2012.

[11] S. Sesia, I. Toufik, and M. Baker, *LTE. The UMTS Long Term Evolution. From Theory To Practice*, John Wiley & Sons Ltd, 2011.

[12] Akyildiz, I. F., Gutierrez-Estevez, D.M. and Reyes, E.C. 'The evolution to 4G cellular systems: LTE-Advanced, ' *Physical Communication*, 3(4) December 2010, pp. 217-244.

[13] NGMN, 'Next Generation Mobile Networks Beyond HSPA & EVDO – A white paper,' *www.ngmn.org*, December 2006.

[14] 3GPP, 'Overview of 3GPP Release 8 V0.2.12,' *Technical Report*, September 2013.

[15] 3GPP, 'Overview of 3GPP Release 9 V0.3.0,' *Technical Report*, September 2013.

[16] 3GPP, 'Overview of 3GPP Release 10 V0.1.10,' *Technical Report*, September 2013.

[17] 3GPP, 'Overview of 3GPP Release 11 V0.1.6,' *Technical Report*, September 2013.

[18] 3GPP, 'Overview of 3GPP Release 12 v.0.1.0,' *Technical Report*, October 2013.

[19] 3GPP, 'Overview of 3GPP Release 13 v.0.0.3,' *Technical Report*, October 2013.

[20] H. Holma and A. Toskala, *LTE for UMTS OFDMA and SC-FDMA Based Radio Access*, John Wiley & Sons Ltd, 2009.

[21] 3GPP, 'Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, Release 12, v.12.2.0, ' *Technical Report*, September 2013.

[22] D. Soldani, M. Li, and R. Cuny, *QoS and QoE Management in UMTS Cellular Systems*, John Wiley & Sons Ltd, 2006.

[23] 3GPP, 'Technical Specification Group Services and System Aspects; Policy and charging control architecture Release 12, v.12.2.0,' *Technical Report*, September 2013.

[24] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2, Release 11, v.11.7.0,' *Technical Report*, September 2013.

[25] 3GPP, 'Technical Specification Group Core Network and Terminals; Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3, Release 12 , v.12.2.0,' *Technical Report*, September 2013.

[26] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification, Release 11, v.11.5.0,' *Technical Report*, September 2013.

[27] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification, Release 11, v.11.2.0,' *Technical Report*, March 2013.

[28] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification, Release 11, v.11.0.0,' *Technical Report*, September 2012.

[29] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification, Release 11, v.11.3.0,' *Technical Report*, June 2013.

[30] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description, Release 11, v.11.1.0,' *Technical Report*, December 2012.

[31] Sabella, D., Caretti, M. and Fantini, R. 'Energy Efficiency Evaluation of State of the Art Packet Scheduling algorithms for LTE, ' *IEEE European Wireless Conference – Sustainable Wireless Technologies*, April 2011, pp. 717-720.

[32] Videv, S. and Haas, H. 'Energy-Efficient Scheduling and Bandwidth–Energy Efficiency Trade-Off with Low Load,' *IEEE International Conference on Communications (ICC)*, June 2011, pp. 1-5.

[33] Frenger, P., Moberg, P., Malmodin, J., Jading, Y. and Gódor, I. 'Reducing Energy Consumption in LTE with Cell DTX,' *IEEE Vehicular Technology Conference (VTC-Spring)*, May 2011, pp. 1-5.

[34] C. Cox, *An Introduction to LTE, LTE-Advanced, SAE and 4G Mobile Communications*, John Wiley & Sons Ltd, 2012.

[35] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, Release 11, v.11.4.0,' *Technical Report*, September 2013.

[36] 3GPP, 'Technical Specification Group Radio Access Network; Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA), Release 7, v.11.4.0,' *Technical Report*, September 2006.

[37] Pokhariyal, A., Pedersen, K. I., Monghal, G., Kovacs, I. Z., Rosa, C., Kolding, T. E. and Mogensen, P.E. 'HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution,' *IEEE Vehicular Technology Conference (VTC-Spring)*, April 2007, pp. 2761 - 2765.

[38] Pedersen, K. I., Frederiksen, F., Kolding, T. E., Lootsma, T. F. and Mogensen, P.E. 'Performance of High-Speed Downlink Packet Access in Coexistence with Dedicated Channels,' *IEEE Transactions on Vehicular Technology*, 56 (3), May 2007, pp.1261–1271.

[39] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures, Release 11, v.11.4.0,' *Technical Report*, September 2013.

[40] Monghal, G. D. A., *Downlink Radio Resource Management for QoS Provisioning in OFDMA systems - with Emphasis on Admission Control and Packet Scheduling*. Ph.D. Thesis, Aalborg University, 2009.

[41] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance, Evolution Towards 3G/UMTS*, John Wiley & Sons Ltd, 2007.

[42] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, The IMT Press Cambridge, 2012.

[43] Comşa, I.-S., Aydin, M., Zhang, S., Kuonen., P. and Wagen, J.F. ' Reinforcement Learning based Radio Resource Scheduling in LTE-Advanced,' *International Conference on Automation & Computing*, September 2011, pp. 219-224.

[44] Comşa, I.-S., Aydin, M., Zhang, S., Kuonen., P. and Wagen, J.F. 'Multi Objective Resource Scheduling in LTE Networks using Reinforcement Learning,' *International Journal of Distributed Systems and Technologies*, 3(2), June 2012, pp. 39-57.

[45] Liu, X., Chong, E. and Shroff, N., ' Opportunistic Transmission Scheduling with Resource-Sharing Constraints in Wireless Networks,' *IEEE Journal on Selected Areas in Communications*, 19 (10) October 2001, pp.2053-2064.

[46] L. Kleinrock*, Queueing Systems, Vol I: Theory*. New York: Wiley, 1975.

[47] Wang, X., Giannakis, G.B. and Marques, A. G., 'A Unified Approach to QoS-Guaranteed Scheduling for Channel-Adaptive Wireless Networks,' Proceedings of IEEE, 95(12) December 2007, pp. 2410-2431.

[48] Wang, X. and Giannakis, G.B., 'A Stochastic Framework for Scheduling in Wireless Packet Access Networks,' *IEEE International Conference on Communications*, June 2007, pp. 4052-4057.

[49] Song, G. and Li (Geoffrey), Y., 'Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks,' *IEEE Communications Magazine*, 43(12) December 2005, pp. 127-134.

[50] Song, G. (2005). *Cross-Layer Resource Allocation and Scheduling in Wireless Multicarier Networks*, Published PhD Thesis, Georgia Institute of Technology.

[51] R.T. Rockafellar, *Convex Analysis*, New Jersey: Princeton University Press, 1970.

[52] R. Irmer, *Radio Access Performance Evaluation Methodology*, Next Generation Mobile Networks Std. V 1.3, 2008.

[53] Kelly, F., 'Charging and Rate Control for Elastic Traffic,' *European Transactions of Telecommunications*, 8(1) January 1997, pp. 33-37.

[54] Lundevall, M., Olin, B., Olsson, J., Wiberg, N., Wanstedt, S., Eriksson, J. and Eng, F., 'Streaming Applications over HSDPA in Mixed Service Scenarios,' *IEEE Vehicular Technology Conference (VTC-Fall)*, 4 September 2004, pp. 841-845.

[55] Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., Whiting, P. and Vijayakumar, R., 'Providing Quality of Service Over a Shared Wireless Link,' *IEEE Communications Magazine*, 39(2) February 2001, pp. 150-154.

[56] Khan, N., Martini, M.G., Bharucha, Z. and Auer, G., 'Opportunistic Packet Loss Fair Scheduling for Delay-Sensitive Applications over LTE Systems,' *IEEE Wireless Communications and Networking Conference*, April 2012, pp. 1456 – 1461.

[57] S. S. Rao, *Engineering Optimization: Theory and Practice*, John Wiley & Sons Ltd, 2009.

[58] R. Rardin , *Optimization in Operations Research*, Upper Saddle River, NJ: Prentice Hall, 1998.

[59] Blum, C. and Roli., A, 'Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison,' *Journal ACM Computing Surveys (CSUR*), 35 (3) September 2003, pp. 268-308.

[60] Bianchi, L., Dorigo, M., Gambardella, L. M. and Gutjahr, W. J., 'A Survey on Metaheuristics for Stochastic Combinatorial Optimization,' *Journal of Natural Computing*, 8 (2) June 2009, pp 239-287.

[61] Gosavi, A., Das, T. K. and Sarkar, S., 'A Simulation-Based Learning Automata Framework for Solving Semi-Markov Decision Problems Under Long-Run Average Reward,' *Journal of IIE Transactions*, 36 June 2003, pp. 557–567.

[62] Shimabukuro, K., Nakamura, M., Ombuki, B.M. and Onaga, K., 'A New Hybrid GA Solution to Combinatorial Optimization Problems and Application to the Multiprocessor Scheduling Problem,' *Artificial Life and Robotics*, 2(2) 1998, pp. 74-79.

[63] Lim, M. H., Jing, T. and Ong, Y. S., 'A Parallel Hybrid GA for Combinatorial Optimization Using Grid Technology,' *The 2003 Congress on Evolutionary Computation (CEC'03)*, 3(8–12) 2003, pp. 1895–1902.

[64] Kwan, R., Leung, C. and Zhang, J., 'Resource Allocation in an LTE Cellular Communication System,' *IEEE International Conference on Communications*, June 2009, pp. 1-5.

[65] Knopp, R. and Humblet, P.A., 'Information Capacity and Power Control in Single-Cell Multiuser Communications,' *IEEE International Conference on Communications*, vol. 1 June 1995, pp. 331-335.

[66] Asadi, A. and Mancuso, V., 'A Survey on Opportunistic Scheduling in Wireless Communications,' *IEEE Communications Surveys and Tutorials*, 15(4) January 2013, pp. 1671 – 1688.

[67] M. L. Pinedo, *Scheduling: Theory, Algorithms and Systems*, Springer, 2012.

[68] Capozzi, F. Piro, G., Grieco, L.A., Boggia, G. and Camarda, P., 'Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey,' *IEEE Communications Surveys and Tutorials*, 15(2) June 2012, pp. 678 - 700.

[69] Nonchev, S. and Valkama, M., 'A New Fairness-Oriented Packet Scheduling Scheme with Reduced Channel Feedback for OFDMA Packet Radio Systems,' *International Journal of Communications, Network and System Sciences, 7,* October 2009, pp. 608-618.

[70] Wunder, G., Zhou, C., Bakker, H.E. and Kaminski, S., 'Throughput Maximization under Rate Requirements for the OFDMA Downlink Channel with Limited Feedback,' *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, August 2007, pp. 1-14.

[71] Pedersen, K.I., Monghal, G, Kovacs, I.Z., Kolding, T.E., Pokhariyal, A., Frederiksen, F. and Mogensen, P., 'Frequency Domain Scheduling for OFDMA with Limited and Noisy Channel Feedback, ' *IEEE Vehicular Technology Conference (VTC-2007 Fall),* September 2007, pp. 1792 – 1796.

[72] Kolehmainen, N., Puttonen, J., Kela, P., Ristaniemi, T., Henttonen, T. and Moisio, M., 'Channel Quality Indication Reporting Schemes for UTRAN Long Term Evolution Downlink, ' *IEEE Vehicular Technology Conference (VTC-2008 Spring)*, May 2008, pp. 2522 – 2526.

[73] Aydin, M.E., Kwan, R., Wu, J. and Zhang, J., 'Multiuser Scheduling on the LTE Downlink with Simulated Annealing,' *IEEE Vehicular Technology Conference (VTC-Spring)*, May 2011, pp. 1-5.

[74] Aydin, M.E., Kwan, R. and Wu, J., 'Multiuser Scheduling on the LTE Downlink with Meta-heuristic Approaches,' *ELSEVIER Physical Communication*, vol. 9 December 2013, pp. 257–265.

[75] Aggarwal, R., Koksal, C.E. and Schniter, P., 'Joint Scheduling and Resource Allocation in OFDMA Downlink Systems Via ACK/NAK Feedback,' *IEEE Transactions on Signal Processing*, 60(6) June 2012, pp. 3217 – 3227.

[76] Ouyang, W., Murugesan, S., Eryilmaz, A. and Shroff, N.B., 'Exploiting Channel Memory for Joint Estimation and Scheduling in Downlink Networks,' *Proceedings of IEEE INFOCOM*, April 2011, pp. 3056 – 3064.

[77] Whittle, P., 'Restless Bandits: Activity Allocation in a Changing World,' *Journal of Applied Probability*, vol. 25 1988, pp. 287-298.

[78] D. P. Bertsekas, *Dynamic Programming and Optimal Control, vol. 1 and 2*, Athena Scientific, Belmount, Massachusetts, 2005.

[79] Schwarz, S., Mehlfuhrer, C. and Rupp, M., 'Throughput Maximizing Multiuser Scheduling with Adjustable Fairness,' *IEEE International Conference on Communications (ICC)*, June 2011, pp. 1-5.

[80] Jain, R., Chiu, D. and Hawe, W., 'A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems, ' Technical Report TR-301, September 1984.

[81] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[82] Proebster, M., Mueller, C.M. and Bakker, H., 'Adaptive fairness control for a proportional fair LTE scheduler,' *IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, September 2010, pp. 1504 – 1509.

[83] Andrews, M., Qian, L., Stolyar, A., 'Optimal Utility based Multi-User Throughput Allocation Subject to Throughput Constraints,' *IEEE Annual Joint Conference Computer and Communications Societies (INFOCOM)*, vol. 4 March 2005, pp. 2415 – 2424.

[84] Tirkkonen, O. and Jantti, R., 'On α-Proportional Fair Packet Scheduling in OFDMA Downlink,' *Annual Allerton Conference on Communication, Control and Computing,* October 2012, pp. 241 – 245.

[85] Li, X., Li, B., Lan, B., Huang, M. and Yu, G., 'Adaptive PF Scheduling Algorithm in LTE Cellular System,' *International Conference on Information and Communication Technology Convergence (ICTC)*, November 2010, pp. 501 – 504.

[86] Wengerter, C., Ohlhorst, J. and von Elbwart, A.G.E., 'Fairness and Throughput Analysis for Generalized Proportional Fair Frequency Scheduling in OFDMA,' *IEEE Vehicular Technology Conference (VTC-Spring)*, vol. 3 June 2005, pp. 1903 – 1907.

[87] Ameigeiras, P., Wigard, J. and Mogensen, P., 'Performance of packet scheduling methods with different degree of fairness in HSDPA,' *IEEE Vehicular Technology Conference (VTC-Fall),* vol. 2 September 2004, pp. 860 – 864.

[88] Mehrjoo, M., Awad, M.K., Dianati, M. and Shen, X., 'Design of Fair Weights for Heterogeneous Traffic Scheduling in Multichannel Wireless Networks,' *IEEE Transactions on Communications*, 58 (10) October 2010, pp. 2892 – 2902.

[89] Yu, W. and Lui, R., 'Dual Methods for Nonconvex Spectrum Optimization of Multicarrier Systems,' IEEE Transactions on Communications, 54 (7) 2006, pp. 1310-1322.

[90] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, 2006.

[91] Hou, I-H. and Chen, C.S., 'Self-organized Resource Allocation in LTE Systems with Weighted Proportional Fairness,' *IEEE International Conference on Communications (ICC)*, June 2012, pp. 5348 – 5353.

[92] Kela, P., Puttonen, J., Kolehmainen, N., Ristaniemi, T., Henttonen, T., Moisio, M., 'Dynamic Packet Scheduling Performance in UTRA Long Term Evolution Downlink,' *International Symposium on Wireless Pervasive Computing (ISWPC),* May 2008, pp. 308 – 313.

[93] Monghal, G., Pedersen, K.I., Kovacs, I.Z. and Mogensen, P.E., 'QoS Oriented Time and Frequency Domain Packet Schedulers for The UTRAN Long Term Evolution,' *IEEE Vehicular Technology Conference (VTC-Spring),* May 2008, pp. 2532 – 2536.

[94] Kolding, T.E., 'QoS-Aware Proportional Fair Packet Scheduling with Required Activity Detection,' *IEEE Vehicular Technology Conference (VTC-Fall)*, September 2006, pp. 1-5.

[95] Catterysse, D. and Van Wassenhove, L., 'A Survey of Algorithms for the Generalized Assignment Problem,' *European Journal of Operational Research*, 60 (3) August 1992, pp. 260–272.

[96] Pitic, R. and Capone, A., 'An Opportunistic Scheduling Scheme with Minimum Data-Rate Guarantees for OFDMA,' *IEEE Wireless Communications and Networking Conference (WCNC)*, April 2008, pp. 1716 – 1721.

[97] Hartmann, C., Vilzmann, R., Schmitt-Nilson, A. and Eberspacher, J., 'Channel aware scheduling for user-individual QoS provisioning in wireless systems,' *IEEE Vehicular Technology Conference (VTC-Fall)*, vol. 2 September 2004, pp. 1009 – 1013.

[98] Ning, X., Ting, Z., Ying, W. and Ping, Z., 'A MC-GMR Scheduler for Shared Data Channel in 3GPP LTE System,' I*EEE Vehicular Technology Conference (VTC-Fall)*, September 2006, pp. 1-5.

[99] Hosein, P.A., 'QoS control for WCDMA high speed packet data,' *International Workshop on Mobile and Wireless Communications Network*, 2002, pp. 169 – 173.

[100] Liu, Q., Wang, X. and Giannakis, G.B., 'A Cross-Layer Scheduling Algorithm With QoS Support in Wireless Networks,' *IEEE Transactions on Vehicular Technology*, 55(3) May 2006, pp. 839 – 847.

[101] Zhang, J., Yuan, D. and Zhang, H., 'Joint Radio Resource Allocation and Scheduling in a Backhaul Constrained Multicell OFDMA Network,' *International Conference on Communication Technology (ICCT)*, September 2011, pp. 47 – 51.

[102] Wang, X., Giannakis, G.B. and Yu, Y., 'Channel-Adaptive Optimal OFDMA Scheduling,' *Annual Conference on Information Sciences and Systems*, March 2007, pp. 536 – 541.

[103] Rasool, J., Hassel, V., De la Kethulle de Ryhove, S. and Øien, G.E., 'Opportunistic Scheduling Policies for Improved Throughput Guarantees in Wireless Networks,' *EURASIP Journal on Wireless Communications and Networking*, vol. 43 July 2011, pp. 1-18.

[104] Rasool, J. and Øien, G.E., 'Maximizing the Throughput Guarantees in Wireless Networks under Imperfect Channel Knowledge,' *IEEE Wireless Communications and Networking Conference (WCNC)*, April 2012, pp. 2225 – 2229.

[105] Ruangchaijatupon, N. and Ji, Y., 'Proportional Fairness with Minimum Rate Guarantee Scheduling in a Multiuser OFDMA Wireless Network,' *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, vol. 7 2009, pp. 1102-1106.

[106] Ameigeiras, P., Wigard, J. and Mogensen, P., 'Performance of the M-LWDF Scheduling Algorithm for Streaming Services in HSDPA, ' *IEEE Vehicular Technology Conference (VTC-Fall)*, vol. 2 September 2004, pp. 999 – 1003.

[107] Stolyar, A.L. and Ramanan, K., 'Largest Weighted Delay First Scheduling: Large Deviation and Optimality, ' *The Annals of Applied Probability*, 11 (1) February 2001, pp. 1-48.

[108] Rhee, J.H., Holtzman, J.M. and Kim, D.-K., 'Scheduling of Real/Non-Real Time Services: Adaptive EXP/PF Algorithm,' *IEEE Semiannual Vehicular Technology Conference (VTC-Spring)*, vol. 1 April 2003, pp. 462 – 466.

[109] Basukala, R., Mohd Ramli, H.A. and Sandrasegaran, K., 'Performance Analysis of EXP/PF and M-LWDF in Downlink 3GPP LTE System,' *First Asian Himalayas International Conference on Internet*, November 2009, pp. 1-5.

[110] Sadiq, B., Madan, R. and Sampath, A., 'Downlink Scheduling for Multiclass Traffic in LTE,' *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, pp. 1-18.

[111] Shakkottai, S., Srikant, R. and Stolyar, A.L (2005). *Pathwise Optimality of the Exponential Scheduling Rule for Wireless Channels*, Published Technical Report, University of Texas at Austin.

[112] Sadiq, B., Baek, S.J. and de Veciana, G., 'Delay-Optimal Opportunistic Scheduling and Approximations: The Log Rule,' *Proceedings of IEEE INFOCOM,* April 2009, pp. 1692 – 1700.

[113] Shakkottai, S. and Srikant, R., 'Scheduling real-time traffic with deadlines over a wireless channel,' *Journal of Wireless Networks*, 8(1) January 2002, pp. 13-26.

[114] Liu, B., Tian, H and Xu, L., 'An Efficient Downlink Packet Scheduling Algorithm for Real Time Traffics in LTE Systems , ' *IEEE Consumer Communications and Networking Conference (CCNC)*, January 2013, pp. 364 – 369.

[115] Bae, S. J., Choi, B.-G. and Chung, M., 'Delay-Aware Packet Scheduling Algorithm for Multiple Traffic Classes in 3GPP LTE System,' *Asia-Pacific Conference on Communications (APCC)*, October 2011, pp. 33 – 37.

[116] Lei, H., Fan, C., Zhang, X. and Yang, D., 'QoS Aware Packet Scheduling Algorithm for OFDMA Systems,' *IEEE Vehicular Technology Conference (VTC-Fall)*, October 2007, pp. 1877 – 1881.

[117] Piro, G., Grieco, L.A., Boggia, G. and Camarda, P., 'A Two-Level Scheduling Algorithm for QoS Support in the Downlink of LTE Cellular Networks,' *European Wireless Conference (EW)*, April 2010, pp. 246 – 253.

[118] Piro, G., Grieco, L.A., Boggia, G., Fortuna, R. and Camarda, P., 'Two-Level Downlink Scheduling for Real-Time Multimedia Services in LTE Networks,' *IEEE Transactions on Multimedia*, 13(5), October 2011, pp. 1052 – 1065.

[119] Ali, S. and Zeeshan, M., 'A Utility Based Resource Allocation Scheme with Delay Scheduler for LTE Service-Class Support,' *IEEE Wireless Communications and Networking Conference (WCNC)*, April 2012, pp. 1450 – 1455.

[120] Sandrasegaran, K., Ramli, H.A.M. and Basukala, R., 'Delay-Prioritized Scheduling (DPS) for Real Time Traffic in 3GPP LTE System,' *IEEE Wireless Communications and Networking Conference (WCNC)*, April 2010, pp. 1-6.

[121] Choi, S., Jun, K., Shin, Y., Kang, S. and Choi, B., 'MAC Scheduling Scheme for VoIP Traffic Service in 3G LTE,' *IEEE Vehicular Technology Conference (VTC-Fall)*, October 2007, pp. 1441 – 1445.

[122] Kim, Y. S., 'An Efficient Scheduling Scheme to Enhance the Capacity of VoIP Services in Evolved UTRA Uplink, ' *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, 2008.

[123] Saha, S. and Quazi, R., 'Priority-Coupling-A Semi-Persistent MAC Scheduling Scheme for VoIP Traffic on 3G LTE,' *International Conference on Telecommunications*, June 2009, pp. 325 – 329.

[124] Wang, L., Kwok, Y., Lau, W., and Lau, V., 'Channel Adaptive Fair Queuing for Scheduling Integrated Voice and Data Services in Multicode CDMA Systems,' Computer Communications, 27(9) June 2004, pp. 809-820.

[125] Fan, Y., Lunden, P., Kuusela, M. and Valkama, M., 'Efficient Semi-Persistent Scheduling for VoIP on EUTRA Downlink, ' *IEEE Vehicular Technology Conference (VTC-Fall)*, September 2008, pp. 1-5.

[126] Mushtaq, M.S., Shahid, A. and Fowler, S., 'QoS-Aware LTE Downlink Scheduler for VoIP with Power Saving,' *IEEE International Conference on Computational Science and Engineering (CSE)*, December 2012, pp. 243 – 250.

[127] Kausar, R., Chen, Y. and Chai, K.K., 'An Intelligent Scheduling Architecture for Mixed Traffic in LTE-Advanced, ' *IEEE International Personal Indoor and Mobile Radio Communications (PIMRC)*, September 2012, pp. 565 – 570.

[128] D. O. Hebb, *The Organization of Behavior, New* York: Wiley, 1949.

[129] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2001.

[130] Eryilmaz, A., Srikant, R. and Perkins, J.R., 'Stable Scheduling Policies for Fading Wireless Channels,' *IEEE/ACM Transactions on Networking*, 13(2) April 2005, pp. 411 – 424.

[131] Liu, S, Ying, L. and Srikant, R., 'Throughput-Optimal Opportunistic Scheduling in the Presence of Flow-Level Dynamics,' *IEEE International Conference on Computer Communications (INFOCOM)*, March 2010, pp. 1 – 9.

[132] A.L. Stolyar (2204). *Maxweight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic*, Annals of Applied Probability, Institute of Mathematical Statistics, 14 (1), pp. 1-53.

[133] Song, G., Li, Y., Cimini, L.J. and Zheng, H., 'Joint Channel-Aware and Queue-Aware Data Scheduling in Multiple Shared Wireless Channels,' *IEEE Wireless Communications and Networking Conference*, vol. 3 March 2004, pp. 1939 – 1944.

[134] Song, G., Li, Y. and Cimini, L.J., 'Joint Channel- and Queue-Aware Scheduling for Multiuser Diversity in Wireless OFDMA Networks,' *IEEE Transactions on Communications*, 57 (7) July 2009, pp. 2109 – 2121.

[135] Kim, H. and de Veciana, G., 'Losing Opportunism: Evaluating Service Integration in an Opportunistic Wireless System,' *IEEE International Conference on Computer Communications*, May 2007, pp. 982 – 990.

[136] Beidokhti, R.K., Hossein, M., Moghaddam, Y. and Chitizadeh, J., 'Adaptive QoS Scheduling in Wireless Cellular Networks,' *Springer Wireless Networks*, 17 (3) April 2011, pp. 701-716.

[137] Comşa, I.-S., Zhang, S., Aydin, M. , Kuonen, P. and Wagen, J., 'A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-advanced Technologies Using Neural Networks,' *IEEE Conference on Local Computer Networks (LCN)*, October 2012, pp. 332 – 335.

[138] Choi, K.W., Jeon, W.S, Jeong, D.G., 'Resource Allocation in OFDMA Wireless Communications Systems Supporting Multimedia Services,' *IEEE/ACM Transactions on Networking*, 17 (3) June 2009, pp. 926 – 935.

[139] Iturralde, M., Ali Yahiya, T., Wei, A. and Beylot, A.-L., 'Performance Study of Multimedia Services Using Virtual Token Mechanism for Resource Allocation in LTE Networks,' *IEEE Vehicular Technology Conference (VTC-Fall)*, September 2011, pp. 1-5.

[140] Bergantinos, G. and Vidal-Puga, J.J., 'Additive Rules in Bankruptcy Problems and Other Related Problems,' *Elsevier Mathematical Social Sciences*, 47(1) January 2004, pp. 87-101.

[141] L. S. Shapley. *A Value for N-Person Game,* Annals of Mathematics Studies, Princeton University Press, vol. 2 1953, pp. 307-317.

[142] Iturralde, M., Ali Yahiya, T., Wei, A. and Beylot, A.-L., 'Resource Allocation Using Shapley Value in LTE Networks,' *IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PMIRC)*, September 2011, pp. 31 – 35.

[143] Iturralde, M., Wei, A., Ali Yahiya, T. and Beylot, A.-L., 'Resource Allocation for Real Time Services Using Cooperative Game Theory and a Virtual Token Mechanism in LTE Networks,' *IEEE Consumer Communications and Networking Conference (CCNC)*, January 2012, pp. 879 – 883.

[144] Monghal, G., Laselva, D., Michaelsen, P.-H. and Wigard, J., 'Dynamic Packet Scheduling for Traffic Mixes of Best Effort and VoIP Users in E-UTRAN Downlink,' *IEEE Vehicular Technology Conference (VTC-Spring)*, May 2010, pp. 1-5.

[145] Chao, I-F. and Chiou, C.-S., 'An Enhanced Proportional Fair Scheduling Algorithm to Maximize QoS Traffic in Downlink OFDMA Systems,' *IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013, pp. 239 – 243.

[146] Chung, Y.-H. and Chang, C.-J., 'A Balanced Resource Scheduling Scheme With Adaptive Priority Thresholds for OFDMA Downlink Systems,' *IEEE Transactions on Vehicular Technology*, 61(3) March 2012, pp. 1276 – 1286.

[147] Ryu, S., Ryu, B., Seo, H. and Shi, M., 'Urgency and Efficiency based Wireless Downlink Packet Scheduling Algorithm in OFDMA System,' *IEEE Vehicular Technology Conference (VTC-Spring)*, vol. 3 June 2005, pp. 1456 – 1462.

[148] Kim, Y., Son, K. and Chong, S., 'QoS Scheduling for Heterogeneous Traffic in OFDMA-Based Wireless Systems,' *IEEE Global Telecommunications Conference (GLOBECOM)*, December 2009, pp. 1-6.

[149] 3GPP, 'Technical Specification Group Radio Access Network; Physical Layer Aspects for Evolved Universal Terrestrial Radio Access (UTRA),' Release 7, v.7.1.0, *Technical Report*, 2006.

[150] 3GPP, 'Technical Specification Group Radio Access Network; High Speed Downlink Packet Access: UE Radio Transmission and Reception (FDD), ' Release 5, v.1.0.0, *Technical Report*, 2005.

[151] 3GPP, 'Technical Specification Group Radio Access Network; Radio Frequency (RF) System Scenarios,' Release 11, v.11.0.0, *Technical Report*, 2012.

[152] 3GPP, 'Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception, ' Release 11, v.9.0.0, *Technical Report*, 2009.

[153] Jakes, W.C., *Microwave Mobile Communication*, Willey, 1975.

[154] International Telecommunication Union, *Guidelines for evaluations of radio transmission technologies for IMT-2000,* ITU ITU-R M.1225, 1997.

[155] Zhang, Y.P., Hwang, Y. 'Measurements of the Characteristics of Indoor Penetration Loss,' *The 44th IEEE Vehicular technology Conference*, 1994.

[156] Piro, G., Grieco, L.A., Boggia, G., Capozzi, F., Camarda, P. 'Simulating LTE Cellular Systems: an Open Source Framework', *IEEE Transactions on Vehicular Technology*, 60 (2) February 2011, pp. 498-513.

[157] A. Gosh, and R . Ratasuk, *Essentials of LTE and LTE-A*, Cambridge University Press.

[158] E. Dahlman, S. Parkvall and J. Sköld, *4G: LTE/LTE-Advanced for Mobile Broadband*, Academic Press, 2011.

[159] Stoffers, M. and Riley, G., 'Comparing the ns-3 Propagation Models,' *The IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, August 2012, pp. 61-67.

[160] Zheng, Y.R., Xiao, C. 'Simulation Models With Correct Statistical Properties for Rayleigh Fading Channels', IEEE Transactions on Communications, 51(6) June 2003, pp. 920-928.

[161] 3rd Generation Partnership Project Technical Specification Groups and Working Group 4 (Radio) Meeting #51, *'Simulation assumptions and parameters for FDD HeNB RF requirements,'* 3GPP TSG WG4, San Francisco, May 4-8 2009.

[162] 3GPP, *'Technical Specification Group Radio Access Network; User Equipment (UE) conformance specification; Radio transmission and reception (FDD); Part 1: Conformance specification (Release 7),'* v.7.2.0, 2006.

[163] Mehlführer, C., Wrulich, M., Colom, I.J, Bosanka, D., Rupp, M. 'Simulating the Long Term Evolution Physical Layer,' *European Signal Processing Conference (EUSIPCO 2009)*, August 2009, pp. 1471- 1478.

[164] Seo, H. and Lee, B.L., 'A Proportional-Fair Power Allocation Scheme for Fair and Efficient Multiuser OFDM Systems,' *IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 6 December 2004, pp. 3737- 3741.

[165] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Second Edition, Prentice Hall, 1998.

[166] Broomhead, D. S. and Lowe, D., 'Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks,' *Technical Report*, 1988.

[167] Park, J. and Sandberg I. W., 'Universal Approximation Using Radial-Basis-Function Networks, ' *Neural Computation,* 3 (2) 1991, pp. 246–257.

[168] Cortes, C. and Vapnik, V., 'Support-Vector Networks,' *Journal of Machine Learning*, 20 (3) September 1995, pp. 273-297.

[169] Hsu, C.-W. and Lin, C.-J., 'A Comparison of Methods for Multiclass Support Vector Machines, ' *IEEE Transactions on Neural Networks*, 13 (2) August 2002, pp. 415 – 425.

[170] Duan, K.-B. and Keerthi, S.S., 'Which Is the Best Multiclass SVM Method? An Empirical Study,' *Proceedings of the 6th International Workshop on Multiple Classifier Systems*, 2005, pp. 278-285.

[171] Lin, S.-W., Lee, Z.-J., Chen, S.-C. and Tseng, T.-Y, 'Parameter Determination of Support Vector Machine and Feature Selection Using Simulated Annealing Approach, ' *Applied Soft Computing for Dynamic Data Mining*, 8 (4) September 2008, pp. 1505-1512.

[172] Zhang, Q., Shan, G., Duan, X. and Zhang, Z., 'Parameters Optimization of Support Vector Machine based on Simulated Annealing and Genetic Algorithm, ' *IEEE International Conference on Robotics and Biomimetics*, December 2009, pp. 1302-1306.

[173] Liu, S. and Jiang, N., 'SVM Parameters Optimization Algorithm and Its Application,' *IEEE International Conference on Mechatronics and Automation*, August 2008, pp. 509-513.

[174] E. E. Osuna and F. Girosi, *Advances in kernel methods*, MIT Press Cambridge, 1998, pp.271-283.

[175] Wang, J. and Wu, X., 'Support Vector Machines Based on K-Means Clustering for Real-Time Business Intelligence Systems, ' *International Journal of Intelligence and Data Mining*, 1 (1) 2005, pp. 54-64.

[176] Baum, E. B. and Haussler, D., 'What Size Net Gives Valid Generalization?, ' *Journal of Neural Computation*, 1 (1) 1989, pp. 151-160.

[177] Zhang, G.P., 'Neural Networks for Classification: A Survey, ' *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 30 (4) November 2000, pp. 451-462.

[178] J. Shawe-Taylor and N. Cristianini, *Kernels Methods for Pattern Analysis*, Cambridge University Press, 2004.

[179] Burges, C.J.C., 'A Tutorial on Support Vector Machines for Pattern Recognition, ' *Data Mining and Knowledge Discovery*, 2 (2) June 1998, pp. 121-167.

[180] M.D. Buhmann, *Radial Basis Functions: Theory and Implementations*, Cambridge University Press, 2003.

[181] Capoyleas, V., Rote, G. and Woeginger, G., 'Geometric Clusterings, ' *Journal of Algorithms*, vol. 12 1991, pp. 341-356.

[182] Kolliopoulos, S. G. and Rao, S., 'A Nearly Linear-Time Approximation Scheme for the Euclidian k-Median Problem, ' *The Seventh Annual European Symposium on Algorithms*, vol. 1643 July 1999, pp. 362-371.

[183] Sharir, M., 'A Near-Linear Algorithm for the Planar 2-Center Problem, ' *Discrete and Computational Geometry*, 18 (2) 1997, pp. 125-134.

[184] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[185] Kanungo, T., Mount, D. M., Netanyahu, N., Piatko, C., Silverman, R. and Wu, A. Y., 'A Local Search Approximation Algorithm for k-Means Clustering,' *Computational Geometry: Theory and Applications*, vol. 28 2004, pp. 89-112.

[186] Lloyd, S.P., 'Least Squares Quantization in PCM, ' *IEEE Transactions on Information Theory*, 28 (2) March 1982, pp. 129-137.

[187] MacQueen, J., 'Some Methods for Classification and Analysis of Multivariate Observations, ' *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 1967, pp. 281-296.

[188] Faber, V., 'Clustering and the Continuous k-Means Algorithm, ' *Los Alamos Science*, vol. 22 1994, pp. 138-144.

[189] Selim, S. Z. and Ismail, M. A., 'K-Means-Type Algorithms: A Generalized Convergence Theorem and Caracterization of Local Optimality, ' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1) January 1984, pp. 81-87.

[190] Charikar, M. and Guha, S., 'Improved combinatorial algorithms for the facility location and k-median problems, ' *40th Annual Symposium on Foundations of Computer Science*, October 1999, pp. 378 - 388.

[191] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P., 'Optimization by Simulated Annealing, ' *Science*, 220 (4598) May 1983, pp. 671-680.

[192] Wenzel, W. and Hamacher, K., 'A Stochastic Tunneling Approach for Global Minimization of Complex Potential Energy Landscapes, ' *Physical Review Letters*, 82 (15) 1999.

[193] Lin, M. and Wawrzynek, J., 'Improving FPGA Placement With Dynamically Adaptive Stochastic Tunneling, ' *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 29 (12) December 2010, pp. 1858 – 1869.

[194] Hamacher, K., 'Adaptation in Stochastic Tunneling Global Optimization of Complex Potential Energy Landscapes, ' *Europhysics Letters*, 74 (6) June 2006, pp. 944-950.

[195] Bentley, J. L., 'Multidimensional Binary Search Trees Used for Associative Searching**,** ' *Magazine Communications of ACM*, 18 (9) September 1975, pp. 509-517.

[196] Tapas, K., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 'An Efficient k-Means Clustering Algorithm: Analysis and Implementation, ' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (7) June 2002, pp. 881 – 892.

[197] A. M. Mood, F. A. Graybill, D. C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill, 1974.

[198] R. J. Aristizabal, *Estimating the Parameters of the Three-Parameter Lognormal Distribution*, Master Thesis, Florida International University, 2012.

[199] Ericsson Mobility Report, 'On the Pulse of the Networked Society,' *Technical Report*, November 2014.

[200] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[201] Karlik, B and Olgac, A.,V., 'Performance Analysis of Various Activation Function in Generalized MLP Architectures of Neural Networks, ' *International Journal of Artificial Intelligence And Expert Systems (IJAE)*, 1(4) 2010, pp. 111-122.

[202] R. Rojas, *Neural Networks. A Systematic Introduction*, Springer, 1996, pp. 151-184.

[203] H. van Hasselt (2011). *Insights in Reinforcement Learning Formal Analysis and Empirical Evaluation of Temporal-Difference Learning Algorithms,* Ph.D. Thesis, University of Utrecht.

[204] Bellman, R., 'A Markovian Decision Process, ' *Journal of Mathematics and Mechanics*, 6 (5) 1957, pp. 679-684.

[205] W. B. Powell, *Approximate Dynamic Programming-Introduction to Markov Decision Processes*, John Wiley and Sons, 2010.

[206] M. L. Puterman, *Markov Decision Processes, Discrete Stochastic, Dynamic Programming*, Wiley Interscience, 1994.

[207] Burnetas, A. N. and Katehakis, M. N., 'Optimal Adaptive Policies for Markov Decision Processes, ' *Mathematics of Operations Research*, 22 (1) 1997, pp. 222-255.

[208] W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New York, 1970, pp. 205-215.

[209] Busunoiu, L., Babuska, R. and Schutter, B. D., 'Multi-Agent Reinforcement Learning: A Survey, ' *Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, December 2006, pp. 527–532.

[210] Busunoiu, L., Babuska, R. and Schutter, B. D., 'A Comprehensive Survey of Multiagent Reinforcement Learning, ' *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews*, 38(2) March 2008, pp. 156-152.

[211] Panait, L. and Luke, S., 'Cooperative Multi-Agent Learning: The State of the Art, ' *Autonomous Agents and Multi-Agent Systems*, 11(3) November 2005, pp. 387-434.

[212] Hu, J., and Wellman, M. P., 'Nash Q-Learning for General-Sum Stochastic Games, ' *Journal of Machine Learning Research*, vol. 4 2003, pp. 1039–1069.

[213] Shoham, Y., Powers, R. and Grenager, T., 'Multi-Agent Reinforcement Learning: a Critical Survey, ' *Technical Report*, Stanford University, May 2003.

[214] Hu, J. and Wellman, M. P., 'Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm, ' *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 242-250.

[215] Chalkiadakis, G. and Boutilier, C., 'Coordination in Multiagent Reinforcement Learning: A Bayesian Approach,' Proceedings of Autonomous Agents and Multi-Agents Systems, July 2003, pp. 1-8.

[216] Claus, C. and Boutilier, C., 'The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems,' *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, 1998, pp.746-752.

[217] Suematsu, N. and Hayashi, A., 'A Multiagent Reinforcement Learning Algorithm using Extended Optimal Response, ' *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, 2002, pp. 370-377.

[218] Littman, M. L., 'Value-Function Reinforcement Learning in Markov Games, ' *Journal of Cognitive Systems Research*, 2 (1), 2001, pp. 55–66.

[219] Guestrin, C., Lagoudakis, M. G. and Parr, R., 'Coordinated Reinforcement Learning,' *Proceedings of the ICML-2002 The Nineteenth International Conference on Machine Learning*, July 2002, pp. 227–234.

[220] Spaan, M. T. J., Vlassis, N. and Groen, F. C. A., 'High Level Coordination of Agents based on Multiagent Markov Decision Processes with Roles,' *IROS'02 Workshop on Cooperative Robotics*, October 2002, pp. 66–73.

[221] Powers, R. and Shoham, Y., 'New Criteria and A New Algorithm for Learning in Multi-agent Systems,' *Proceedings of Advances in Neural Information Processing Systems*, vol. 17 December 2004, pp. 1089–1096.

[222] Weiβ, G., 'Distributed Reinforcement Learning,' *Robotics and Autonomous Systems*, vol. 15 1995, pp.135-142.

[223] Rogova, G., Scott, P. and Lolett, C., 'Distributed reinforcement learning for sequential decision making,' *Proceedings of the Fifth International Conference on Information Fusion*, vol. 2 July 2002, pp. 1263 – 1268.

[224] Lauer, M. and Riedmiller, M., 'An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems,' *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 535-542.

[225] Lauer, M. and Riedmiller, M., 'Reinforcement Learning for Stochastic Cooperative Multi-Agent Systems,' *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 3 2004, pp. 1516-1517.

[226] R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.

[227] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics, 1951, pp. 400–407.

[228] C. J. C. H. Watkins, *Learning from Delayed Rewards*, Ph.D. Thesis, Cambridge University, 1989.

[229] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic Programming*, Athena Scientific, 1996

[230] Tsitsiklis, J. N. and Van Roy, B., 'An Analysis of Temporal-Difference Learning with Function Approximation, ' *IEEE Transactions on Automatic Control*, 42 (5) 1997, 674–690.

[231] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer New York, 2006.

[232] Wilson, D. R. and Martinez, T. R., 'The General Inefficiency of Batch Training for Gradient Descent Learning,' *Neural Networks*, 16(10) 2003 pp.1429–1451.

[233] Watkins, C. J. C. H. and Dayan, P., '{Q}-Learning', *Machine Learning Journal- Special Issue on Reinforcement Learning*, 8(3/4) May 1992.

[234] H. P. van Hasselt, 'Double Q-Learning, ' *Advances in Neural Information Processing Systems*, the MIT Press, vol. 23 2010.

[235] Rummery, G. and Niranjan, M., *On-line Q-learning using Connectionist Systems*, University of Cambridge, Engineering Department, Technical Report. no.166, 1994.

[236] Singh, S. P., Jaakkola, T., Littman, M. L. and Szepesvari, C., 'Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms,' *Machine Learning*, 38 (3) 2000, pp. 287-308.

[237] Wiering, M.A., 'QV(lambda)-learning: A New On-policy Reinforcement Learning Algorithm,' *Proceedings of the 7th European Workshop on Reinforcement Learning*, 2005, pp. 17-18.

[238] Wiering, M.A. and van Hasselt, H, 'The QV Family Compared to Other Reinforcement Learning Algorithms,' *Proceedings of IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, 2009, pp. 101 - 108.

[239] Van Hasselt, H., and Wiering, M., 'Using Continuous Action Spaces to Solve Discrete Problems,' *Proceedings of the International Joint Conference on Neural Networks*, 2009, pp. 1149 - 1156.

[240] Van Hasselt, H. and Wiering, M. 'Reinforcement Learning in Continuous Action Spaces,' *Proceedings of IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL07)*, 2007, pp. 272-279.

[241] Jain, R., Chiu, D.M. and Hawe, W., 'A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System, ' *Research Report*, DEC-TR-301, 1984.

[242] 3GPP2 C.R1002-0, 'CDMA2000 Evaluation Methodology, ' Revision 0, 2004.

[243] Senarath, G., Tong, W. et. al., 'Multi-hop Relay System Evaluation Methodology - Channel Model and Performance Metric, ' IEEE 802.16 Broadband Wireless Access Working Group, 2007.

[244] IEEE P 802.2, '802.20 Evaluation Criteria, ' Version 1.0, IEEE P 802.20 PD-09, 23 September 2005.

[245] Comşa, I.-S., Aydin, M., Zhang, S., Kuonen, P., Wagen, J.-F, and Lu, Y., 'Scheduling Policies Based on Dynamic Throughput and Fairness Tradeoff Control in LTE-A Networks,' *IEEE Conference on Local Computer Networks (LCN)*, September 2014, pp. 418 - 421.

[246] Comşa, I.-S., Zhang, S., Aydin, M., Chen, J., Kuonen, P., and Wagen, J.-F, 'Adaptive Proportional Fair Parameterization Based LTE Scheduling Using Continuous Actor-Critic Reinforcement Learning,' *IEEE Global Communication Conference (GLOBECOM)*, September 2014, pp. 4387 - 4393.

[247] Granado, J. M., Vega, M. A., Pérez, R., Sánchez, J. M. and Gómez, J. A., 'Using FPGAs to Implement Artificial Neural Networks,' *IEEE International Conference on Electronics, Circuits and Systems*, December 2006, pp.934-937.

[248] Ferrer, D., Gonzalez, R. , Fleitas, R. , Acle, J.P. , and Canetti, R., 'NeuroFPGA-Implementing Artificial Neural Networks on Programmable Logic Devices,' *Proceedings on Europe Conference and Exhibition in Design, Automation and Test,* vol. 3 February 2004, pp. 218-223.

[249] C. D. de Souza, Alisson and Fernandes, Marcelo A. C., 'Parallel Fixed Point Implementation of a Radial Basis Function Network in an FPGA,' *Sensors*, vol. 14(10) September 2014, pp. 18223-18243.

[250] Joy Vasantha Rani, S.P., Kanagasabapathy, P. and Suganthi, L., ' Field Programmable Gate Array Based Floating Point Hardware Design of Recursive k-means Clustering Algorithm for Radial Basis Function Neural Network, ' *International Journal of Intelligent Systems Technologies and Applications*, vol. 6(1/2) January 2009, pp. 61-76.